COMPUTATIONAL MODELING OF GENE EXPRESSION FROM REGULATORY SEQUENCES

BY

MD. ABUL HASSAN SAMEE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Associate Professor Saurabh Sinha, Chair
Professor Chandra Chekuri
Assistant Professor Jian Ma
Professor Stanislav Shvartsman, Princeton University

Abstract

Regulation of gene expression is an important early step in controlling every biological process that underlies the function of living organisms. Even though gene expression may be regulated in several stages, the modulation occurs mostly at the primary stage known as “transcription”. Teasing out the details of transcriptional regulation is therefore a core focus of biological research. Transcriptional regulation of gene expression is dictated by regulatory DNA sequences, often called *cis*-regulatory modules (CRM; also known as “enhancer”), that contain specific binding sites for regulatory proteins (transcription factors, TF). The assembly of TFs bound on a CRM drives the desired expression level of the gene associated with the CRM. As the abundance of TFs vary across different cell types, the expression level of the gene, also termed as the “readout” of the CRM, varies accordingly and results in the aforementioned control over biological processes. The rules, collectively known as the “*cis*-regulatory logic”, to predict gene expression level given information about CRMs and TFs, however, are unclear. Decades of experimental studies have hypothesized mechanisms about parts of this regulatory process (e.g., about the influence of TF-TF interactions), but a comprehensive study of *cis*-regulatory logic is feasible only through computational models. The subject of this thesis is to develop mechanistic models of gene expression from regulatory sequences and use the models to understand such details of the system that are difficult to assess experimentally.

The first part of this thesis develops a model that integrates the regulatory effect of signaling pathways with that of sequence-bound TFs to understand the expression pattern of a gene from its CRM. Given the various types of molecular interactions that the model needs to capture, it is both complex in structure and rich in the number of parameters. Similarly complex models commonly used in other disciplines, from signaling networks to climatology, have been shown to fit many distinct parameterizations that are equally consistent with data but might represent disparate mechanistic hypotheses. Whether this is also the case for models of *cis*-regulation has never been investigated, with the standard practice in this realm being to report a single or a few best-fit models. We demonstrate here – taking the *Drosophila ind* gene as an example – that gene expression modeling from *cis*-regulatory sequences may suffer from incomplete and even incorrect conclusions if one adheres to this current practice. We construct an ensemble of models by systematically exploring the entire parameter space and leveraging both wild-type data and various perturbation experiments, and make statistical inferences from the ensemble about detail regulatory mechanisms of *ind*. Years of genetic experiments have put forth an assortment of hypotheses about *ind* regulation. We use our modeling approach to show how a mechanism involving MAPK induced attenuation in the DNA binding affinity of Capicua and the use of low-affinity Dorsal binding sites may provide a coherent explanation of *ind* regulation. Also, we quantitatively predict and experimentally validate the role of the “pioneer factor” Zelda in activating *ind*. Finally, we discuss disparate hypotheses that are supported by our ensemble of models and will need future experimentation for a complete understanding of *ind* regulation.

The second part of this thesis addresses a fundamental goal of computational biology, namely that of modeling a gene’s expression from its intergenic locus and trans-regulatory context. Owing to the distributed nature of *cis*-regulatory information and the poorly understood mechanisms that integrate

such information, gene locus modeling is a more challenging task than modeling individual enhancers. Here we report the first quantitative model of a gene's expression pattern as a function of its locus. We model the expression readout of a locus in two tiers: 1) combinatorial regulation by transcription factors bound to each enhancer is predicted by a thermodynamics-based model and 2) independent contributions from multiple enhancers are linearly combined to fit the gene expression pattern. The model does not require any prior knowledge about enhancers contributing toward a gene's expression. We demonstrate that the model captures the complex multi-domain expression patterns of anterior-posterior patterning genes in the early *Drosophila* embryo. Altogether, we model the expression patterns of 27 genes; these include several gap genes, pair-rule genes, and anterior, posterior, trunk, and terminal genes. We find that the model-selected enhancers for each gene overlap strongly with its experimentally characterized enhancers. Our findings also suggest the presence of sequence-segments in the locus that would contribute ectopic expression patterns and hence were "shut down" by the model. We applied our model to identify the transcription factors responsible for forming the stripe boundaries of the studied genes. The resulting network of regulatory interactions exhibits a high level of agreement with known regulatory influences on the target genes. Finally, we analyzed whether and why our assumption of enhancer independence was necessary for the genes we studied. We found a deterioration of expression when binding sites in one enhancer were allowed to influence the readout of another enhancer. Thus, interference between enhancer activities was a possible factor necessitating enhancer independence in our model.

The third part of this thesis applies the aforementioned models to two novel datasets. The first dataset was created by fusing two well-studied CRMs of the *even-skipped* (*eve*) gene in *Drosophila*. The fused constructs differ in the way the CRMs' orientation, order, and intervening spacing are varied. Interestingly, the two constituent CRMs regulate *eve* expression by using the same TFs, although binding affinities (i.e., strength) of the repressor sites in the two CRMs are different – an observation that has been implicated to help the CRMs drive expression in two distinct domains (each domain consists of two stripes of *eve*) when they act in their endogenous context. However, the fact that these two CRMs harbor sites for the same TFs makes it difficult to predict the readouts of the constructs in our dataset. In particular, readouts of these constructs show some subtle aspects that essentially challenge the conventional models of information integration from sequences and suggest that a different mechanism may be necessary to explain these observations. Our modeling of this novel dataset suggests that the conventional assumption that relatively short DNA sequences, *e.g.*, CRMs, do not comprise smaller "independent" regulatory sequences may not be true – since the lengths of the fused constructs are comparable to typical CRMs and their readouts can be modeled by assuming the existence of smaller independent regulatory segments. The second dataset modeled in this part of the thesis features five genes that control the growth and patterning of wing in *Drosophila*. Notably, ours is the first attempt to link regulatory sequences and the related molecular details to the growth and scaling of an organ. In course of fitting this dataset, we identify the important regulatory role of a TF called Scalloped (Sd) and speculate on Sd's role in assuring that the expression domains of the studied genes scale with wing growth. We also use our models to identify novel regulatory sequences of these genes and to answer several questions that were left open in the experimental studies that attempted first to understand the *cis*-regulatory logic for these genes.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Saurabh Sinha, for his constant support, guidance, and mentorship throughout my PhD studies. Saurabh not only stayed watchful that I do not lose my focus as a graduate student, but also inspired me continuously to aim at and plan for higher achievements. I am deeply indebted to him for the way he taught me to become uncompromising about soundness of arguments and quality of presentations. Above all, he was by my side at the times of my personal and family crises.

Every member of my Dissertation Committee has played a role much beyond what one would expect from a thesis committee member. They have shown utmost affection and care, and encouraged me to shoot higher, although they knew very well about my weaknesses. I would like to thank Dr. Chandra Chekuri for guiding me from my first day at UIUC, and I will always remember how he pushed me to "learn to ask the right question". Dr. Stanislav Shvartsman, despite our physical distance, has interacted with me and suggested me for improvements as intensely as Saurabh did. I am grateful to him for the many discussions we had online and also for arranging my visits to Princeton. I cannot thank Dr. Jian Ma enough for the hours he spent advising me on planning my career and offering me every privilege during my postdoctoral job search.

I will thank here all my teachers and mentors – from my elementary school to UIUC – for their patience and love in building me up. My deepest gratitude goes to Bangladesh University of Engineering and Technology for providing me with university education for literally no cost – the significance of which I realized only after coming to the USA.

This thesis would not have been possible without the love and support from my parents, my wife, and my beloved daughter. I do not know the appropriate words to thank them. I am also indebted to my elder sister and my brother in law for taking care of many of those issues which I was responsible for.

My friends and my group-mates in the Sinha Lab have been my constant sources of all sorts of support. I do not want to name any one specifically; even only their first names would take an entire chapter in this write up.

Finally, I thank the Almighty for blessing me with these wonderful six years of my life.

Table of Contents

List of Figures

Chapter 1
Introduction

Cellular and organismal processes depend critically on the establishment of complex gene expression patterns at precise times and spatial locations (Davidson 2006). Mis-regulation is increasingly implicated in a broad range of disease states, and changes in gene expression underlie morphological differences between species (Maurano, Humbert et al. 2012). The information for directing complex gene expression patterns is encoded in regulatory DNA sequences, also known as *cis*-regulatory modules (CRM) or enhancers. Genes exhibiting complex expression patterns (in temporal and/or spatial domain) are typically regulated by multiple CRMs; each CRM regulates the corresponding gene's expression in a distinct spatio-temporal context. The regulatory control of a CRM is conferred by an assembly of transcription factor (TF) molecules that bind to their cognate sites on the CRM and control the activity of the molecular complexes (referred to as the basal transcription machinery, BTM) that transcribe the corresponding gene (Ptashne 2002) (Fig. 1.1).

Decades of genetic experiments have revealed that assembly of TFs on a CRM is not sufficient to achieve regulatory control. Precise regulatory control is achieved by constraining the strength and/or the arrangement of binding sites on a CRM, and it depends on how TFs interact with each other, with other molecular species (e.g., molecules in signaling pathways), and with the BTM. Essentially, regulatory control of gene expression is the outcome of a "transcriptional program" that follows a "*cis*-regulatory logic" (Levo and Segal 2014, Weingarten-Gabbay and Segal 2014) (Fig. 1.2). Given the central role of gene expression regulation in many biological processes, a predictive and quantitative understanding of the *cis*-regulatory logic is desirable. Such an understanding would allow us to go beyond merely identifying the TFs and CRMs that are involved in a biological process, and would replace the existing qualitative and phenomenological descriptions with a mechanistic view of the process that integrates the components that are involved into realistic mechanistic models (Fig. 1.3). Models of gene expression from CRMs – aptly known as sequence-to-expression models – thus have direct applications in comparative genomics, identification of functional SNPs in GWAS and eQTL analyses, and design of sequences in synthetic biology and gene therapy (Levo and Segal 2014). This thesis aims at building computational models of this genre and applying them on quantitative data of gene expression during early embryonic development of *Drosophila melanogaster* (fruitfly).

As transcriptional regulation across different organisms uses the same types of molecules, which interact according to the universal laws of physical chemistry, the basic rules of our models apply broadly. Indeed, different mechanistic aspects modeled here were shown functional in bacteria (Bintu, Buchler et al. 2005), yeast (Kaplan, Moore et al. 2009), flies (Segal, Raveh-Sadka et al. 2008) and mammals (Sinha, Adler et al. 2008). Recent thrust to generate large amounts of data on how transcription factors bind DNA sequences and how these binding events produce expression patterns (Roy, Ernst et al. 2010) has paved the way to build the class of quantitative models we opt for. Furthermore, high-throughput assays (e.g., STARR-Seq) (Sharon, Kalma et al. 2012, Arnold, Gerlach et al. 2013) are cataloguing the expression patterns driven by almost any sequence in the genome – enabling tests to probe the extent to which these models may explain *cis*-regulatory logic and to identify the necessary modifications toward building more realistic

models. In the rest of this chapter, we introduce four such projects where we have worked with the aforementioned goal to develop and apply models of transcriptional regulation from regulatory sequences.

1.1 An Ensemble Approach in Sequence-to-Expression Modeling: Application in Eliciting the *Cis*-Regulation of a Neuro-Ectodermal Gene in *Drosophila*

The *cis*-regulatory logic that drives the early expression of the *intermediate neuroblasts defective* (*ind*) gene in *Drosophila* has remained elusive to date (Stathopoulos and Levine 2005, Garcia and Stathopoulos 2011). The *ind* gene is a classic example of how a precise expression domain results from combinatorial interaction between graded and uniformly expressed TFs and their interplay with activating input from the MAPK signaling pathway. Our objective in this work was to develop a mechanistically rich model to capture the possible interactions between the regulators of *ind*, and explore the entire parameter space for models that may be considered reasonably good agreements with the data on *ind* regulation. The latter goal is justified from several recent works that argue that conventional approaches of identifying a single model by likelihood-maximization techniques may fail to capture the most plausible mechanistic hypothesis underlying the data. As such, we go beyond seeking the model that provides the best agreement with data to computing model ensembles and identifying all possible explanations of the data (Janssens, Hou et al. 2006, Parker, White et al. 2011, Kim, Martinez et al. 2013). This also required us to deal with a second challenge, namely to exploit the identified models to reveal diverse yet plausible mechanistic explanations of the available data, which in turn can led to systematic design of future experiments.

More specifically, we performed here a systematic exploration – following the full factorial sampling strategy – of the entire parameter space of our proposed *cis*-regulatory model for *ind* and identified many different parameter combinations that are equally good in explaining the wild-type expression data. Next, we used these numerous models to predict the gene's expression pattern under various *cis*- and *trans*-perturbations for which data is available. Only models that survive this test of prediction were retained for further analyses. The resulting ensemble of models unanimously explains several previously reported observations and provides novel insights about the gene's regulation. Importantly, by examining the differences among these models, either in parameter values or in terms of their prediction on new perturbation situations, we were able to identify the remaining uncertainties in our understanding of the underlying *cis*-regulatory logic and could provide suggestions for future experimental investigations so as to obtain a complete and predictive quantitative model of the gene's regulation.

Several biological insights from this study are as follows. First, we show that a novel mechanism involving ERK-dependent relief of *ind* repression by Capicua (Cic), proposed recently in studies of EGFR signaling in cultured human cells, likely plays an important role in this gene's regulation. We explain how Cic repression can set the dorsal boundary of *ind* in the blastoderm stage embryo, even though Cic nuclear concentration is constant across the D/V axis – this has been one of the least understood aspects of *ind* regulation and is now an elegant example of how extracellular signaling may confer transcriptional precision by acting at the sequence level. Secondly, we offer a comprehensive picture of Dl and Zld binding sites in the *ind* CRM and their relative roles, with an explanation of why a recent study (Garcia and

Stathopoulos 2011) found no significant change in *ind* expression upon mutating a high-affinity Dl site in the *ind* CRM, even though prior evidence suggests Dl as a major activator of this gene. In particular, we demonstrated the importance of the weak binding sites of Dl in activating *ind* – a question that is of much interest at present due to an increasing appreciation of the role of weak TF sites in evolution and developmental precision. Along the same lines, we also predicted and experimentally verified a reduction in *ind* expression upon Zld site mutagenesis. Finally, based on our ensemble models we proposed a set of experiments which can reduce the remaining uncertainty in *ind* regulation.

1.2 Multi-Tier Models of Gene Expression from Intergenic sequences or "Loci"

While state of the art sequence-to-expression models focus on gene expression modeling from experimentally characterized CRMs, we took here the first step toward the broader goal of predicting the expression readout of arbitrary genomic sequences. This first step involves understanding the mechanisms of how the sequence of a gene's locus may regulate the entire expression pattern of the gene. We decided here not to use any *a priori* knowledge about the CRMs of the gene, since an important goal here was to understand the role of an entire locus in driving the expression pattern of a gene despite the fact that a minimal set of CRMs is often sufficient to capture all aspects of the expression pattern. The first challenge here was therefore to deal with the unknown locations of CRMs in a locus. A second major challenge in this work was to model the mechanisms that integrate outputs from distinct CRMs into the endogenous gene expression. Moreover, when the locus harbors multiple CRMs with similar readouts, as has been suggested by the discovery of "shadow CRMs" (Hong, Hendrix et al. 2008), we attempted to explain how to take into account the contributions from all of them.

To address the above challenges, we developed here GEMSTAT-GL (GEMSTAT-Gene Locus), a quantitative model of a gene's expression pattern as a function of the sequence of its entire locus. We focused on the expression of the genes *even-skipped*, *hairy*, *runt*, and *giant* in the developmental stage following the maternal to zygotic transition in early *Drosophila* embryos. In this stage, each of these genes has a multi stripe expression pattern along the A/P axis that is known to be controlled by multiple enhancers within the locus, and is thus an ideal test for our model. We started with our statistical thermodynamics-based model GEMSTAT that was shown previously to accurately model ~40 enhancers involved in A/P patterning (He, Samee et al. 2010, Samee and Sinha 2013). Following conventional wisdom (Howard, Ingham et al. 1988, Howard and Struhl 1990, Ishihara, Sato et al. 2008, Perry, Boettiger et al. 2011), we then framed our working hypothesis that the expression readout of an entire gene locus is two tiered – sites within each enhancer act together to produce that enhancer's contribution, and contributions from multiple enhancers are combined via yet unknown rules to produce the gene's expression pattern. Pursuing this hypothesis, we showed that a gene's expression can be modeled as a weighted sum of expression driven by several enhancers within its locus, where each enhancer's output is predicted by the thermodynamics-based GEMSTAT model. From the intergenic locus of each gene, our model automatically selects a handful of segments that together generate the gene's expression. In order to demonstrate the broader applicability of GEMSTAT-GL, we next used it to model the expression patterns of 23 additional genes in early *Drosophila* embryo. From the intergenic locus of each gene, our model automatically selected one or a handful of segments that together generated the gene's expression. The selected segments were

found to overlap CRMs known to regulate the gene, even though the model was not informed about these CRMs.

An immediate practical benefit of our model is the automatic discovery of candidate CRMs in the locus, along with accurate assignments of regulatory activity to each CRM. This goes one step beyond our previous work (Kazemian, Blatti et al. 2010) where CRMs were annotated based on their pattern generating potential. The new method ensures that activities of multiple CRMs in the locus can be aggregated to match the gene's expression profile. Also, since GEMSTAT-GL allows model parameters to be trained simultaneously with the discovery of CRMs in a gene's locus, the assignment of regulatory activity to CRMs is empirically more accurate than those reported in (Kazemian, Blatti et al. 2010). We performed *in silico* knock-downs of TFs and identified the TFs responsible for the formation of stripe boundaries in A/P expression patterns of these genes. The resulting network of regulatory interactions exhibits a very high level of agreement with known regulatory influences on the target genes, illustrating the potential of the model-based approach for unraveling regulatory networks. We also developed a method to investigate whether and why the assumed independence of CRMs was necessary in our model. We found that interaction or "cross-talk" (Kirstein, Sanz et al. 1996, Yao, Phin et al. 2008, Prazak, Fujioka et al. 2010) between the CRMs of a gene is detrimental to our model's fits to the gene's expression data, and identified cases where specific binding sites in one CRM that may interfere with another CRM's readout. This suggests that in these cases the independence of CRM contributions is necessary for proper modeling of gene expression. We also investigated whether and how the intergenic sequence outside these selected segments contributes to the gene's expression. Our findings suggest the presence of sequence segments in the locus that would exert an irreconcilable impact on the gene's expression pattern and thus were required to be explicitly "shut down" by the model, presumably reflecting a similar phenomenon *in vivo.*

1.3 Thermodynamic Modeling of Fused Enhancer Constructs to Reveal Novel Mechanism of Transcriptional *Cis*-Regulation

State of the art research have shown the adequacy of *cis*-regulatory models that translate the strength of individual binding sites and the overall site content of a CRM into its expression readout (Janssens, Hou et al. 2006, Gertz, Siggia et al. 2009, Fakhouri, Ay et al. 2010, He, Samee et al. 2010, Parker, White et al. 2011, Kwasnieski, Mogno et al. 2012, Kim, Martinez et al. 2013, Samee and Sinha 2013). On the other hand, recent computational models of intergenic loci (Samee and Sinha 2014) have also pointed to the possibility that the readout of an entire locus may not model the expression pattern of a gene; rather a locus comprises multiple active regulatory sequences that act independently and whose individual readouts are aggregated, while the rest of the locus remains inactive and does not affect gene expression. This contrast is expected given the change in the length of the modeled DNA (i.e., from CRMs which are ~1Kb in length to loci which are ~20-90 Kb in length) and our increasing understanding of the role of chromatin remodeling and epigenetic modifications that delineate active regulatory sequences in a locus (Weingarten-Gabbay and Segal 2014). Our objective here was to probe for the length scale that gives rise to this dichotomy. In other words, we asked "what is the length scale beyond which a sequence has to be

treated as a locus, rather than a CRM"? To this end, we asked in this work an alternative and more tractable question: "can CRM-length sequences comprise active and inactive sequences?"

We chose here a recent dataset of expression patterns driven by sequences that are created by fusing two well-studied CRMs regulating the *even-skipped* gene in Drosophila. We fit the statistical thermodynamics based GEMSTAT model on this data under various hypothesis on how the entire site content of a sequence may generate an observed expression pattern. This showed us the extent to which these different models may explain the data. We also fit the GEMSTAT-GL model and assessed whether a delineation of active and inactive regulatory sequences may explain the data better. Indeed, we find that the readouts of the sequence constructs in this data set are best explained as a linear superposition of independent readouts from distinct sub-elements of the construct. We find that each of the two component CRMs includes sub-elements that may be capable of driving expression in an approximately correct spatial location but to a lower level than the component CRM itself. This points to a novel mechanism where even CRM-length sequences may comprise distinct active and inactive sequences.

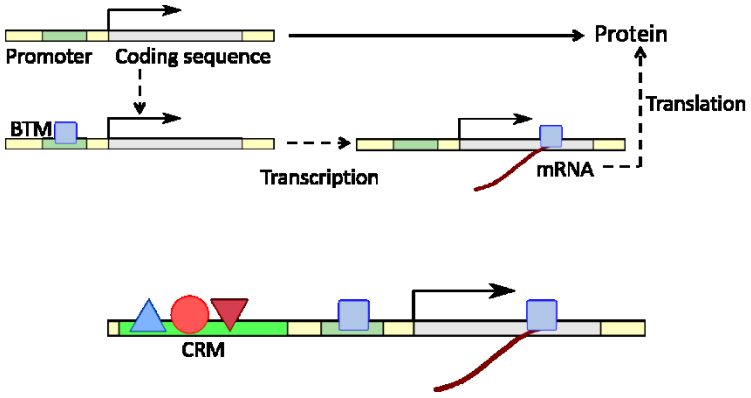
1.4 Computational Modeling of Gene Expression in the Developing Wing Imaginal Disc in *Drosophila*

State of the art models of organ development, *e.g.*, those of wing development in *Drosophila*, focus on morphology of the organ but do not elicit molecular explanations at the level of regulatory sequences (Yin, Xiao et al. 2013, Zhang, Alber et al. 2013). Therefore, the comprehensiveness of the existing knowledge about *cis*-regulation of the genes involved in organ development and growth has never been assessed, although an understanding of the molecular details at the sequence level is the ultimate goal in every developmental investigation (Davidson 2006, Davidson 2010). We attempted in this study to link the expression patterns and the dynamic changes of the *Drosophila* wing developmental genes to their *cis*-regulatory logic.

We studied here a dataset that comprises the set of five genes whose expression initiates the patterning of the *Drosophila* wing imaginal disc in the anterior-dorsal compartment. Of particular biological significance in this compartment is the development of the vein L2 from the cells that express the gene *knirps* (*kni*). We fit this dataset using the GEMSTAT model and use the fit models to investigate the details of formation of the expression patterns of the studied genes. In particular, we have applied the models to complement the known experimental observations on these genes and quantitatively assess various hypotheses of how these genes are expressed. Our analysis suggests important and novel roles for a gene called *scalloped* (*sd*) which has drawn attention for years as an important gene in wing development, but was never demonstrated quantitatively to play the hypothesized roles (Guss, Benson et al. 2013). We also quantitatively predict the effects of knocking-down these genes and show how our model predictions remain consistent with known genetic experimental results. Finally, we use the model to search for additional CRMs of the genes modeled in this work.

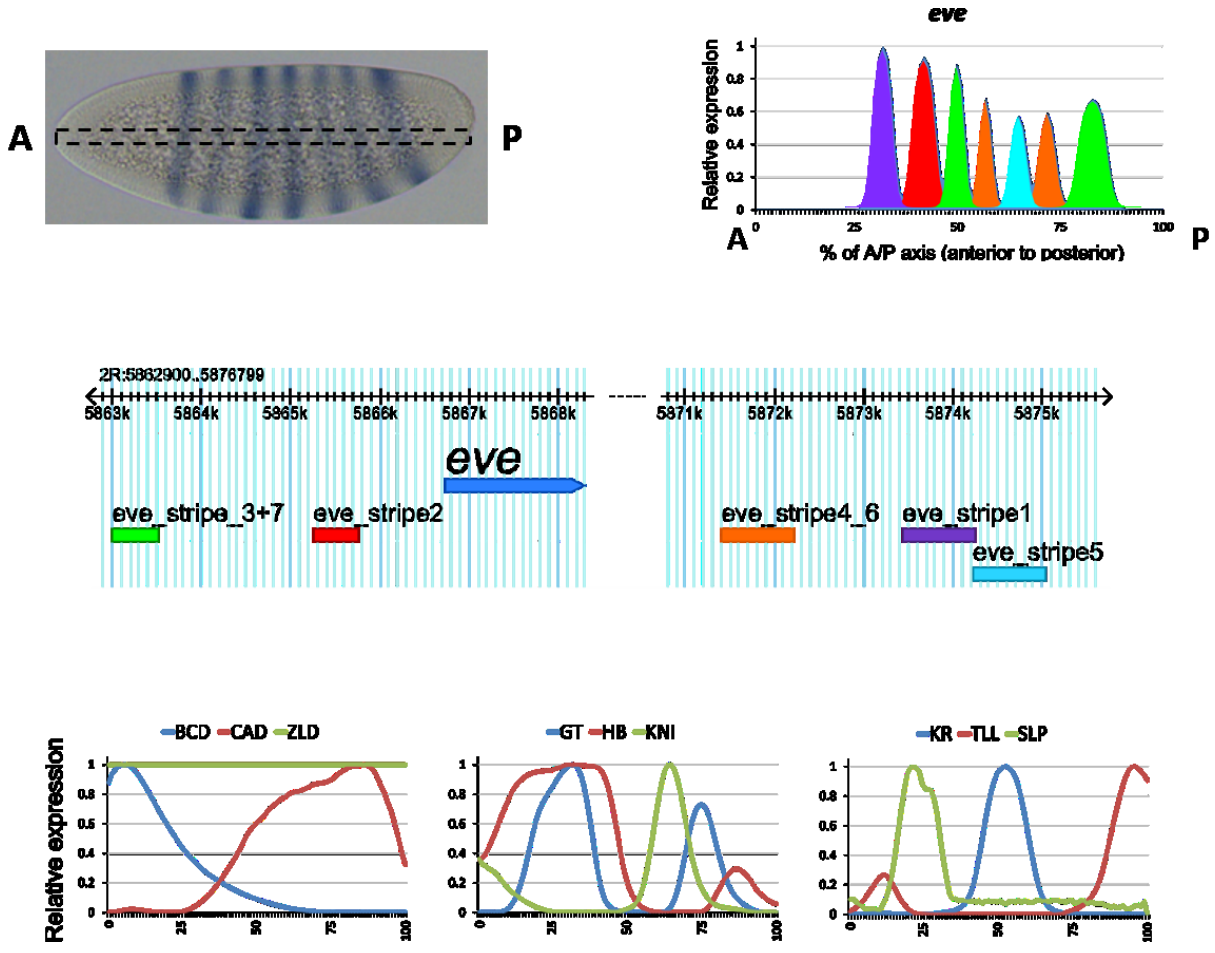
1.5 Figures

Figure 1.1: The regulation of gene transcription through transcription factor (TF) molecule recruitment at specific regulatory sequences (CRM).



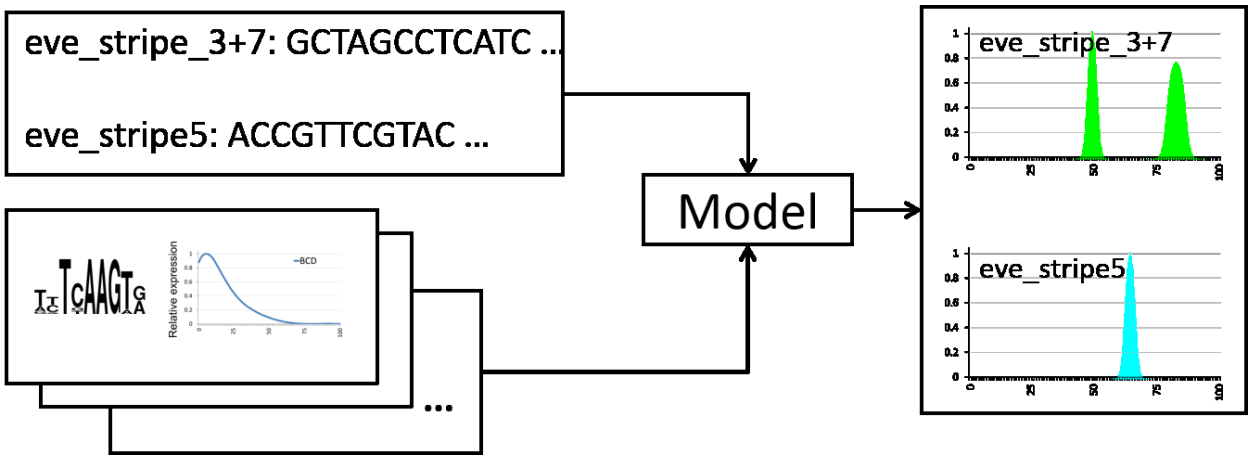
Top panel: overview of steps in protein production from gene coding sequence. Bottom panel: TF assembly at CRM regulates the recruitment of Basal Transcription Machinery (BTM) at gene promoter, which in turn transcribes the gene.

Figure 1.2: Cis-Regulatory Control of the Drosophila even-skipped *gene*



Top panel: A blastoderm-stage embryo and quantitative data extracted from it. The seven stripes show the seven distinct expression domains of the gene even-skipped. Middle panel: the five distinct CRMs of eve. Each sequence is color coded with the stripe that it regulates. Bottom panel: the five CRMs contain the binding sites for nine regulatory TFs and the logic to map their spatial gradients into the seven stripes of eve.

Figure 1.3: Framework of a sequence-to-expression model: inputs are the sequences, the sequence specificities of TFs, and the expression profiles of TFs which are mapped by the model to the expression readout of the sequences.



Chapter 2
Background

In this chapter, we will discuss some basics of modeling transcriptional *cis*-regulation, parameter optimization algorithms, and the strategy of parameter estimation through an ensemble approach. We will also discuss a goodness-of-fit score for scoring and ranking models of transcriptional *cis*-regulation. Some sections in this chapter are summarized from our published works, namely the sections on thermodynamic models (He, Samee et al. 2010, Samee and Sinha 2013), design of scoring function (Samee and Sinha 2013), and comparison of global and local parameter estimation techniques (Suleimenov, Ay et al. 2013).

2.1 Modeling Transcription Factor Binding Specificity

Transcription factors (TF) are regulatory proteins that identify specific DNA sequences (typically short, 6-15 bps in length) as their binding target (Kazemian, Pham et al. 2013). In order to model the specificity of TF binding, one starts with a set of experimentally identified (and also aligned, if necessary) sequences where the TF has been found to bind *in vivo*. Then a matrix called position weight matrix (PWM) is constructed such that the element *W(b, i)* of the matrix specifies the score a sequence *S* should receive if the base *b* (*b* = A, C, G, or T) occurs at position *i* of *S*. One can then compute a score, *Score(S|W)* as $\sum_{b,i} W(b,i)S(b,i)$, where *S(b,i) = 1* if the *i*-th base in *S* is *b*, and *0* otherwise. The goal is therefore to model the PWM such that instances of the TFs binding sites are scored higher than random sequences.

A common approach to this problem is to construct a probabilistic score by defining *W(b, i) = log(F(b, i))* where $F(b, i) = \sum_k S_k(b,i)/N$, *N* is the number of instances of the TFs binding site (Stormo and Zhao 2010). As such, it is being assumed that base occurrences at every position *i* follow a multinomial distribution with parameters *F(b, i)* and we are using the maximum likelihood estimates for these parameters at each position. Note that, the logarithm is taken in the definition of *W(b,i)* so that the additivity in the definition of *Score(S|W)* may be retained. To make the score discriminative against random sequences, Schneider et al. (Schneider, Stormo et al. 1986) proposed to model *W(b, i)* as the log-odds score *log(F(b, i)/p(b))* where *p(b)* denotes base occurrences in the genomic background. In particular, the information content as each position *i* computed from the PWM entries *W(b,i)* denotes the Kullback-Leibler divergence of the genomic background base distribution from the multinomial distribution at *i*.

Berg and von Hippel, and later Stormo and his colleagues, showed an interesting connection to the above definition of *W(b,i)* with the energy of TF-DNA binding at position *i* (Berg and von Hippel 1987, Stormo and Fields 1998) – which has made PWMs readily incorporable into biophysical models of TF-DNA binding and transcriptional regulation. In particular, if each position *i* contributes independently to the free energy of TF-DNA binding, and the entries *W(b,i)* in the PWM denotes the energy contribution of base *b* at position *i*, then the total binding energy *E* at a given site *S* is $\sum_{b,i} W(b,i)S(b,i)$, where *S(b,i) = 1* if the *i*-th base in *S* is *b*, and *0* otherwise. It was shown in (Stormo and Fields 1998) that if the collection of sites used to build the PWM are selected based on their binding affinities, then it corresponds to the sites being selected from a Boltzmann distribution, and the energy contribution at a given position *i* by base *b* equates

log(F(b, i)/p(b)). See Fig. 2.1 for an example of a PWM and its motif "logo" that shows the KL-divergence (with respect to a uniform genomic background) at each base's location.

2.2 Thermodynamic Modeling of Transcriptional Regulation

We discuss here a model called GEMSTAT (Gene Expression Modeling Based on Statistical Thermodynamics) which we proposed first in (He, Samee et al. 2010, Samee and Sinha 2013) for modeling gene expression from regulatory sequences. GEMSTAT estimates the probability of gene expression from the ensemble of all possible configurations of bound TFs and BTM. To this end GEMSTAT computes a "statistical weight" Z_c for each configuration c from the energies of the protein-DNA, BTM-DNA, protein-BTM, and protein-protein interactions in c (Shea and Ackers 1985). Below we elaborate on the computation of Z_c (see Fig. 2.2).

2.2.1 Statistical weight of a configuration

For a configuration c, the expression for Z_c has terms reflecting binding of TFs to their cognate sites and those reflecting TF-TF interactions. If c is a BTM-bound configuration, then Z_c will have additional terms reflecting TF-BTM interactions and the binding of BTM to promoter (Figure 1). Through these four types of terms, Z_c captures the energy of various binding and interaction events occurring in c, where the ground state for computing the energies is a configuration where no TF or the BTM is bound. We explain below how different types of binding and interaction events are accommodated in the formulation of Z_c.

2.2.2 TF-DNA binding

For a given site S, the binding of a TF f at S contributes a statistical weight of $q_{f,S} = K_{f,S}[f]$ to Z_c. Here $K_{f,S}$ is the equilibrium constant of the DNA-binding reaction between f and S, and $[f]$ is the concentration of f. Let S^f_{max} denote the strongest binding site of f and $K(S^f_{max})$ denote the association constant of the TF-DNA binding reaction between f and S^f_{max}. Then we can re-write $K_{f,S}$ as $K(S^f_{max})\exp(-\beta\Delta E_{f,S})$, where $\beta = {}^1/_{k_BT}$, k_B is the Boltzmann constant, T is the temperature, and $\Delta E_{f,S}$ denotes the "mismatch energy" of the site S relative to S^f_{max} for f. According to the theory of Berg and von Hippel (Berg and von Hippel 1987), we can estimate $\exp(-\beta\Delta E_{f,S})$ from $\exp(-LLR(f,S) + LLR(f,S^f_{max}))$, where $LLR(f,\cdot)$ is the log likelihood ratio score of a site, computed based on the known position weight matrix (PWM) of f and the background nucleotide distribution (Stormo 2000).

The concentration $[f]$ of the TF f is in arbitrary units and essentially can be re-written as $v[f]_{rel}$ where $[f]_{rel}$ is the concentration of f relative to some unknown reference value v. The expression for $q_{f,S}$ then becomes:

$$q_{f,S} = K(S^f_{max})v[f]_{rel}\exp(LLR(f,S) - LLR(f,S^f_{max}))$$

where both $K(S^f_{max})$ and v are unknown quantities. We take their product $K(S^f_{max})v$ as a free parameter in our model and refer to it as the "DNA-binding parameter" for the TF f. We note that, the estimated values of different TFs' DNA-binding parameters in our model are not biochemically comparable since this

parameter represents a product of a biochemical parameter (i.e., $K(S^f_{max})$) and an unknown reference value (i.e., v). We also note that, owing to this formulation, we can fit our model using the *relative* levels of mRNA and TF expression.

In case the site S is a signaling pathway response element and the signaling activity is known to attenuate the DNA binding affinity of f, then using the concentration of a chemical species $Sgnl$ whose spatial distribution correlates with the signal's level of activity, we model a modification of $q_{f,S}$ as follows.

$$q_{f,S,Sgnl} = K(S^f_{max})v[f]_{rel}\exp(LLR(f,S) - LLR(f,S^f_{max}) - \varphi([Sgnl]))$$

We used $\varphi([\mathrm{dpERK}]) = C \times [\mathrm{dpERK}]$ in this study, where C is a free parameter, to model an attenuation of the DNA binding affinity of Cic under the influence of ERK.

2.2.3 TF-TF interaction

If two TFs f_1 and f_2 interact when bound to closely located sites (with no other TF bound between them), as opposed to one TF binding independently of the other (Shea and Ackers 1985), then for each such instance of f_1 and f_2 bound in a configuration c, the statistical weight Z_c includes an extra multiplicative term ω_{f_1,f_2}. This term represents the energy of interaction between f_1 and f_2. As such, $\omega_{f_1,f_2} > 1$ or < 1 depending on whether the interaction between f_1 and f_2 enhance or diminish their occupancy in those closely located sites. Note that, f_1 and f_2 may denote the same TF.

2.2.4 TF-BTM interaction

We assume each TF f to impart on the BTM a "transcriptional effect", which essentially represents the energy of interaction between f and the BTM. As such, for each instance of f binding to one of its cognate sites in a BTM-bound configuration c, the statistical weight Z_c includes an extra multiplicative term α_f. If f facilitates the recruitment of BTM, then f is a transcriptional activator and $\alpha_f > 1$. Similarly, $\alpha_f < 1$ if f is a transcriptional repressor.

2.2.5 BTM-promoter binding

We include a parameter q_{BTM} in Z_c for every BTM-bound configuration c to capture the energy of BTM binding to promoter.

Considering all the possible binding events occurring in a configuration c, we then write the term Z_c as:

$$Z_c = \left(\prod_S \left({q_{f,S}}^{\sigma_{f,S}} \prod_{\substack{S' < S \text{ and} \\ N(S',S)=0}} {\omega_{f,g}}^{\sigma_{f,S} \times \sigma_{g,S'}} \right) \right) \left(q_{BTM} \prod_S {\alpha_f}^{\sigma_{f,S}} \right)^{\sigma_{BTM}}$$

Where

- Sites in the enhancer are ordered according to their location in a scan of the enhancer (either 5' to 3' or 3' to 5'),
- $\sigma_{f,S}$ is an indicator variable (0/1) to denote where TF f binds to site S,
- σ_{BTM} is an indicator variable (0/1) to denote whether c is a BTM-bound configuration, and
- $N(S', S)$ denotes the number of TF-occupied sites located between two specific sites S' and S where $S' < S$.

As mentioned in the Results section, we ultimately compute the probability of BTM-bound configurations, i.e.,

$$\text{P(bound BTM)} = \frac{Z_{\text{bound}}}{Z_{\text{unbound}} + Z_{\text{bound}}},$$

where the denominator $Z_{\text{unbound}} + Z_{\text{bound}}$ equates Z, the partition function. An efficient computation of the partition function involves application of dynamic programming and the relevant formulations are given in detail in (He, Samee et al. 2010).

2.3 Identifying Putative Binding Sites for Transcription Factors

The above description of our thermodynamic modeling framework assumes that binding sites in a sequence have been annotated beforehand. In theory, such an annotation is not necessary since every *k*-bp window in a CRM is a potential binding target for a TF, where *k* denotes the length of the TF's motif (represented by a position weight matrix, PWM). However, most of these targets are weak-affinity sites and presumably do not represent stable and functional binding. On the other hand, regarding every *k*-mer as a putative binding target will increase computational overhead in model optimization. As such, we apply a thresholding scheme (described below) to mark only the relatively strong binding sites of a TF and discard any site that fails to satisfy the threshold. In particular, to annotate a TF's binding sites in a CRM, we first compute the log likelihood ratio (LLR) score of each *k*-bp window in the CRM, where *k* denotes the length of the TF's motif and the two likelihoods in the ratio are computed from the PWM and a uniform background distribution. A window is then annotated as a binding site for the TF if the window's LLR score is at least half the LLR score of the TF's optimal site. In our experience of working with other datasets of *Drosophila* developmental gene regulation, this scheme can capture the TF's experimentally annotated sites in the given CRM. The motif PWMs used in this model were all collected from the FlyFactorSurvey database (Zhu, Christensen et al. 2011).

2.4 Designing Goodness of Fit Scores

An important part of modeling gene expression is the score used to assess "goodness of fit" between model predictions and data. Typically, this score is also used as the objective function of the model-training algorithm. The popular choice among the handful of quantitative models published previously has been either the Pearson correlation coefficient ("CC") (Zinzen, Senger et al. 2006, Segal, Raveh-Sadka et al. 2008, He, Samee et al. 2010) or the sum of squared errors ("SSE") (Janssens, Hou et al. 2006, Gertz,

Siggia et al. 2009, Fakhouri, Ay et al. 2010). Both scores are easy to compute and a natural choice when the data being modeled is a vector of expression values. However, neither CC nor SSE is sensitive to the shape as well as magnitude of expression, and can lead to counter-intuitive assessments of goodness-of-fit. CC is invariant to scaling or shifts of the compared variables, as a result of which a model's prediction can receive a perfect score (CC=1) even though biological intuition would dictate otherwise (see Figure 2A). On the other hand, the SSE can lead to counter-intuitive assessments because it is insensitive to the sign of the difference between the actual and the predicted expression. An example scenario is shown in Figure 2B. In some cases, the sensitivity of SSE to the magnitude of difference between real and predicted expression levels can lead to undesirable situations, as shown in Fig. 2.3. These shortcomings of the CC and SSE scores were also discussed in our previous work (Kazemian, Blatti et al. 2010), where we proposed a new scheme called "Pattern Generating Potential" (PGP) for comparing gene expression profiles, and argued that it addresses those shortcomings.

The PGP score was designed for binary expression profiles, i.e., where the data specifies cells or nuclei where the gene is expressed and where it is not. This was a reasonable assumption in that work, where the expression data being modeled comprised the one-dimensional enhancer readouts we mentioned in the previous subsection. These expression profiles were of low resolution and carried little information beyond the binary information about expression domains. However, the expression profiles derived from BDTNP are of higher resolution and include quantitative information on mRNA levels. Therefore, we built upon the PGP score to define the "weighted pattern generating potential' (w-PGP) score, which serves as the goodness-of-fit function in this study. As explained in the following paragraph, the w-PGP score extends the core ideas of the PGP so as to handle quantitative (i.e., continuous valued, rather than binary) expression profiles.

The PGP score rewards predicted expression in domains of expression and penalizes predicted expression in domains of non-expression. The final score is based on these reward and penalty terms. To be more specific, consider a real expression profile that specifies axial positions or "bins" where the gene is expressed (denoted by the set E), and bins where the gene is not expressed (set $\overline{E}$). The predicted expression profile is a vector $\boldsymbol{p}$ of expression levels (on a scale of 0 to 1), one prediction p_i for each bin i. The reward term of PGP is then computed by considering all bins in E, and averaging the predicted expression p_i in these bins. Similarly, the penalty term is computed by examining all bins in $\overline{E}$ and averaging the predicted expression p_i in these bins. The w-PGP score does not operate on binary expression profiles, and in place of E and $\overline{E}$ the expression profile has an expression value r_i for every bin i, and the predicted profile assigns a value p_i to each of these bins. The amount of correctly predicted expression in any bin can then be defined as $min(p_i,r_i)$, and in the w-PGP scheme the contribution of this bin to the reward term is defined as r_i x $min(p_i,r_i)$. In other words, the amount of correctly predicted expression is *weighted* by the real expression level in that bin. Bins with greater expression levels contribute more to the reward term. Similarly, the contribution of any bin to the penalty term is defined as $(1-r_i)$ x $abs(p_i-r_i)$. The factor $abs(p_i-r_i)$, which represents the amount of false prediction (either over- or under-prediction) is *weighted* by the extent of non-expression $(1-r_i)$ in that bin. Note that w-PGP is a linear combination of reward and penalty terms such that its value ranges between 0 (bad prediction) and 1 (perfect prediction).

2.5 Parameter Optimization: Local and Global Search Algorithms

In order to identify the optimal model parameters, one needs first to identify a suitable optimization algorithm. In one of our studies on this issue (Suleimenov, Ay et al. 2013), we have compared several such algorithms and found that given the current quality (resolution) of gene expression data, local search algorithms or an engineered combination thereof might be a suitable strategy.

Local estimation techniques generally require either the calculation or approximation of the objective function's derivative or a comparison of the objective function at multiple different parameter values. Some common local parameter estimation techniques include the Conjugate Gradient method, Newton's method, and the Nelder-Mead algorithm. In two earlier thermodynamic-based modeling studies, local parameter estimation techniques were used (Segal, Raveh-Sadka et al. 2008, He, Samee et al. 2010). Segal and colleagues employed the conjugate gradient ascent and the Nelder-Mead simplex methods in an alternating fashion. In a later study, Sinha and colleagues applied the quasi-Newton and the Nelder-Mead simplex methods in an alternating fashion on a slightly different model using the same data set. The major drawback of using such local parameter estimation techniques is that they may lead to the discovery of local, not global, minima of the objective function, producing misleading results. This problem can be minimized if the modeler has prior knowledge of where the global optimal parameter values lie, allowing for an initial guess close to the global minima and thus quick convergence by a local algorithm. In many real-world settings, including the modeling of transcription, almost no prior knowledge of key parameter values is available, however. Thus, the only way to overcome the problem of getting stuck at local minima is to run the algorithm multiple times, with different initial parameter values, a so called multi-start strategy. Multiple starts with different initial parameter values were implemented in both of the studies mentioned above. However, this method is not very efficient, and it has been shown that for highly nonlinear inverse problems, local parameter estimation strategies cannot find the correct parameters even when run with an extremely large number (>300) of starting points (Moles, Mendes et al. 2003).

Global parameter estimation techniques offer another path to finding global minima of a model when no a priori information on parameters is available. However, with these techniques, obtaining the global minima for a nonlinear model is still very difficult, depending on the parameter landscape, and computationally expensive. Both deterministic and stochastic global techniques have been developed. Deterministic methods such as the branch-and-bound and interval optimization methods are more reliable, but they are computationally very expensive and impractical for many nonlinear problems. In contrast, stochastic methods such as genetic algorithms, simulated annealing, and evolutionary strategies can more quickly find the location of global minima. These methods move through parameter space with some stochasticity to avoid getting "stuck" at local minima. Global convergence has been proven for evolutionary algorithms provided a unique global minimum exists, and a nonzero probability of reaching the neighborhood of that minimum from any initial starting population in a single evolutionary time step. Unfortunately, global optimality cannot be guaranteed for all problems due to their probabilistic nature, therefore multiple runs are advised.

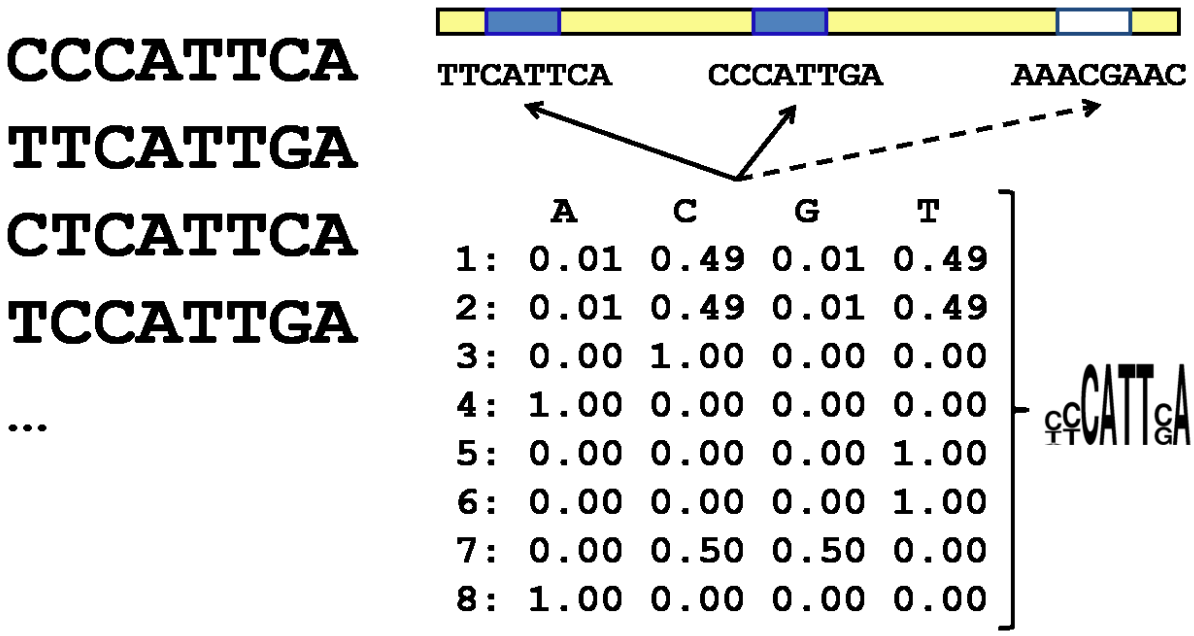
In the realm of stochastic techniques, evolutionary strategies (ES) have shown excellent performance in many studies of continuous models with large sets of estimated parameters. These strategies are inspired

by biological evolution, and include features such as crossover, mutation and selection. Specific techniques vary in the number of offspring and parents, whether mutation rates, recombination, or crossover are considered, and the selection strategy applied. In each ES, the “fittest individuals” (the parameter sets with lowest value for the objective function), have a higher probability of surviving to the next “generation”. Along with selection criteria, the use of mutation rates, recombination, and crossover allow populations of parameter sets to leave local minima to reach a global minimum. In the same study (Suleimenov, Ay et al. 2013), we used a version of an evolutionary strategy, the so called covariance matrix adaptation–evolutionary strategy (CMA–ES). Our choice was motivated by the documented success of CMA–ES algorithms over other global parameter estimation techniques on benchmark problems.

Nonlinear models, such as thermodynamic-based models of gene regulation, are known to be ill-posed, that is, either they do not have a solution, the solution is not unique, or the solution does not depend continuously on the data. In these problems the parameter space is usually unknown and the complexity of the parameter estimation problem grows exponentially as the number of parameters increase. For this reason, it may be worth employing computationally expensive global parameter estimation techniques. However, in several recent thermodynamic-based modeling studies, local strategies rather than global counterparts were employed to achieve computational efficiency. It is not clear what the potential trade-off was in these cases: the fits obtained were judged adequate, but perhaps the use of a local strategy had a significant effect on fits with possible misleading biological interpretations. To better understand the possible trade-offs in using these diverse parameter estimation approaches, it is necessary to compare performance on the same dataset, using the same models. Here, we test the performance of the CMA–ES global and QN/NMS local methods, with respect to fitting parameter values in thermodynamic-based models of gene regulation. We compared these algorithms using both synthetic and experimental gene expression data. In designing the synthetic datasets, we used a score called “the derivative score” that we designed to score expression patterns generated from real enhancers but random parameters to score and identify patterns for their non-randomness. The score computes the derivative of the computed expression pattern at each data point and takes the average of the absolute values of these derivatives to quantify the amount of non-randomness of the pattern generated by the parameter.

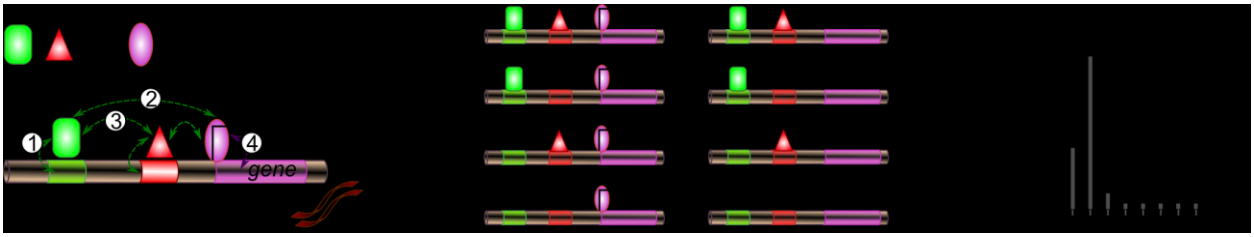
2.6 Figures

Figure 2.1: Overview of Position Weight Matrix Computation



Left: sequences of binding sites, right: the matrix where each entry corresponds to the frequency of that corresponding base and that position and the corresponding motif logo computed from KD-divergence between the base frequencies and a uniform background distribution at each postion.

Figure 2.2: Overview of the thermodynamics based GEMSTAT model



(A) GEMSTAT models the major components and their interactions involved in transcriptional regulation: the CRM (DNA sequence), transcription factors (TFs), and the basal transcriptional machinery (BTM). TFs bind at their cognate sites in the CRM and the BTM binds at the promoter. The mRNA expression level is determined by strength of TF-DNA interactions (at the binding sites) and TF-BTM interactions. Different possible interactions are shown with arrows. (B) GEMSTAT assumes that the system is at thermodynamic equilibrium. An exponential number of possible configurations of bound TFs and the BTM may occur in equilibrium. Shown are the eight possible configurations corresponding to the example shown in A. (C) GEMSTAT assumes the mRNA expression level is proportional to the equilibrium probability of the BTM binding at the promoter. Under standard statistical mechanical assumptions, equilibrium probabilities of configurations follow Boltzmann distribution. Shown is a hypothetical probability distribution for the configurations shown in B. The probability of BTM binding at promoter is computed from the probabilities of all BTM-bound configurations (i.e., configurations c1—c4).

Figure 2.3: The Weighted Pattern Generating Potential Score

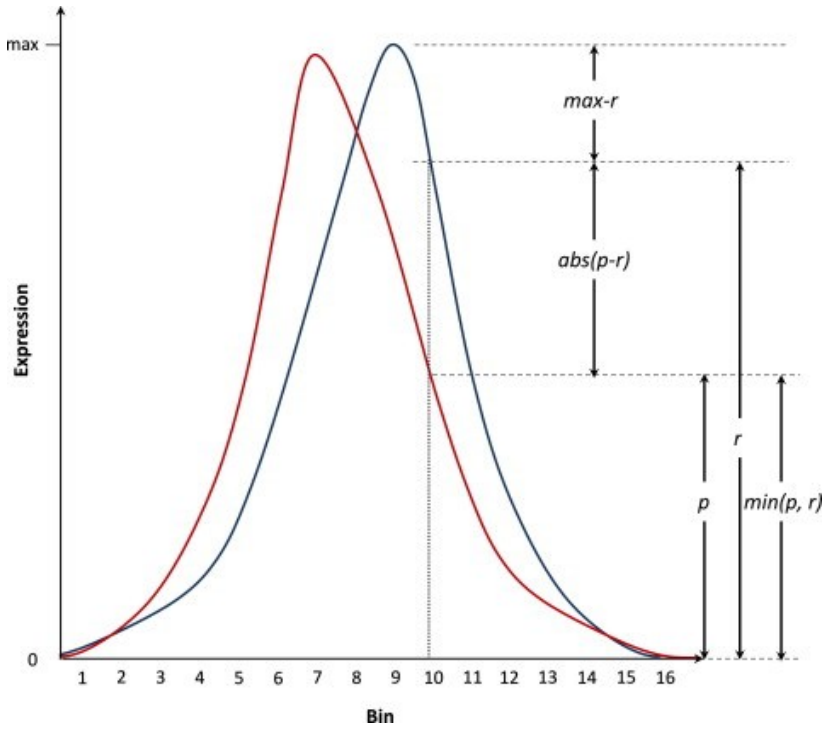


Illustration of the various terms involved in computing the w-PGP score. The blue and the red curves represent the actual and the predicted expression profiles, respectively. The profiles have 16 data points (bins) each. Terms of the w-PGP formulation are illustrated with respect to data point 10.

Chapter 3
An Ensemble Approach to Predict Gene Expression from Regulatory Sequences

3.1 Introduction

The integration of various regulatory inputs at genomic regulatory sequences known as *cis*-regulatory modules (CRMs) is an important early step in regulating eukaryotic gene expression. CRMs harbor specific sites where regulatory proteins (transcription factors, TF) bind and regulate transcription of a target gene (Yanez-Cuna, Kvon et al. 2013). As TF abundance varies across different cell types, CRM-controlled expression level of the gene, also termed as the "readout" of the CRM, varies accordingly. However, the rules to predict gene expression level given information about CRMs and TFs, sometimes collectively known as the "*cis*-regulatory logic", are unclear (Yanez-Cuna, Kvon et al. 2013, Weingarten-Gabbay and Segal 2014). Understanding the general principles of this logic and its implementation in specific biological contexts is one of the grand challenges in biology today. Genetic experiments often facilitate such investigations by eliciting isolated pieces of information, e.g., how "knocking-down" a TF affects a gene's expression pattern. However, the ultimate goal is a unifying picture (model) that explains the available assortment of experimental results both qualitatively and quantitatively, suggests experiments to improve upon the current model, and is capable of predicting the gene's expression pattern upon alterations of its *cis*- or *trans*-determinants. This motivates the development of quantitative models that map regulatory sequence to associated gene expression levels in given cellular contexts (Ay and Arnosti 2011). We refer to such models here as "sequence-to-expression" models. Note that the formulation and capabilities of such models are distinct from the numerous existing attempts to reconstruct transcriptional regulatory networks that are based on modeling one gene's expression as a function of other genes' expression levels (Perkins, Jaeger et al. 2006).

Transcriptional regulatory mechanisms are known to be complex (Yanez-Cuna, Kvon et al. 2013), and additional levels of complexity continue to be discovered (Levo and Segal 2014). Sequence-to-expression models must capture essential elements of this mechanistic complexity, and at the same time make simplifications that allow those models to be trained on the generally sparse datasets available. One of the simplest such modeling paradigms uses sequence-specific descriptions of binding site affinities (Stormo 2000) to model variable site affinities and transcription rates within an equilibrium thermodynamics framework (Shea and Ackers 1985). Thermodynamic models of this genre are arguably more realistic, yet do not use more parameters, than models where all sites of a TF are assumed to have the same affinity (Zinzen, Senger et al. 2006, Zinzen and Papatsenko 2007, Papatsenko and Levine 2008, Fakhouri, Ay et al. 2010) or only two types of affinity, viz. "strong" and "weak" (Bintu, Buchler et al. 2005, Gertz, Siggia et al. 2009, Parker, White et al. 2011, White, Parker et al. 2012). We previously reported one such sequence-to-expression model called GEMSTAT, and used it for modeling ~40 CRMs involved in anterior-posterior patterning in early stages of *Drosophila* development (He, Samee et al. 2010). Equilibrium thermodynamics is an established framework today for modeling transcriptional *cis*-regulation (Ay and Arnosti 2011, Sherman and Cohen 2012). As mentioned above, GEMSTAT incorporates

sequence-specific TF-DNA interaction energy in the thermodynamic framework. It focuses on molecular interaction energies involving TF proteins, DNA and the basal transcriptional machinery, and uses statistical thermodynamics to map combinations of interactions to the probability of finding the CRM in a configuration supporting transcription. Changes in sequence and/or TF concentrations translate to changes in these interactions and consequently to changes in the rate of transcription.

Typically, quantitative models of gene expression include free parameters whose values are learned during model training so as to obtain the best fit between model and data, after which this optimal model may be used to predict the effects of *cis* or *trans* perturbations. For instance, GEMSTAT and other models resort to parameter fitting to estimate equilibrium constants of TF-DNA binding and the effects of a DNA-bound TF on gene expression. These parameters are TF-specific (1-3 parameters per TF (Zinzen, Senger et al. 2006, Segal, Raveh-Sadka et al. 2008, Fakhouri, Ay et al. 2010, He, Samee et al. 2010, White, Parker et al. 2012, Kim, Martinez et al. 2013)) and, given that typical CRMs are controlled by a handful of TFs (Berman, Nibu et al. 2002, Berman, Pfeiffer et al. 2004, He, Samee et al. 2010, Kazemian, Blatti et al. 2010, Samee and Sinha 2013), even simple models may have a considerable number of free parameters. A consequence of the high dimensionality of parameter space is that the model may make predictions consistent with available data at many distinct parameter settings. Sequence-to-expression models have been successfully used in the literature, but to what extent their parameters are constrained by data is unclear. It is entirely possible that in the typical scenario there are multiple optima in the parameter space. In such cases, any of the multiple optima or near-optima in the parameter space may represent the underlying mechanisms, within the assumed modeling framework.

The above considerations argue for going beyond the common practice of seeking the single best fitting model, and instead exploring the entire parameter space for models exhibiting strong agreement with data. Although some related studies have found widely different parameter assignments to fit a dataset (Granek and Clarke 2005, Zinzen and Papatsenko 2007), the standard practice is to report one or at most a few best models that result after iterative improvements upon a small sample of initial random guesses (Janssens, Hou et al. 2006, Segal, Raveh-Sadka et al. 2008, He, Samee et al. 2010, Parker, White et al. 2011, Kim, Martinez et al. 2013). In this work, we performed a systematic exploration of the entire parameter space of the GEMSTAT model for a specific developmental gene – *intermediate neuroblasts defective* (*ind*) – and identified many different parameter combinations that are equally good in explaining the wild-type expression data. We used this diverse ensemble of models to predict the gene's expression pattern under various *cis*- and *trans*-perturbations, and related the predictions to experimental data.

The subject of our analysis is the homeobox transcription factor gene *ind* – a transcriptional target of the Dorsal (Dl) morphogen – that plays an important role in dorso-ventral (D/V) patterning of the ventral nerve cord in *Drosophila.* The embryonic spatio-temporal pattern of this gene has been accurately characterized (Weiss, Von Ohlen et al. 1998) and a well-delineated CRM driving its neuro-ectodermal expression pattern is known (Stathopoulos and Levine 2005). Prior work has revealed or suggested identities of its major regulatory inputs, including activation by the morphogen Dl (Hong, Hendrix et al. 2008) and the ubiquitously expressed Zelda (Zld) (Nien, Liang et al. 2011), and repression by Snail (Sna) (Cowden and Levine 2003), Ventral neuroblasts defective (Vnd) (McDonald, Holbrook et al. 1998, Weiss, Von Ohlen et al. 1998), and Capicua (Cic) (Lim, Samper et al. 2013). Binding specificities (motifs) of these

TFs are also available (Zhu, Christensen et al. 2011). Furthermore, a selection of genetic perturbations and the resulting changes in *ind* expression have been reported in the literature (Stathopoulos and Levine 2005, Garcia and Stathopoulos 2011, Lim, Samper et al. 2013). In this work, we use GEMSTAT to formally integrate this information about *ind* regulation into a quantitative description of combinatorial action by these known TFs and develop a predictive model to study the effects of *cis*- and *trans*-perturbations on *ind* expression.

The parameters of our model include two special parameters (one pertaining to TF-DNA binding and one to the TF's effect on transcription rate) for each of the five hypothesized TFs, a parameter for Dl-Zld cooperativity, and a parameter reflecting the baseline transcriptional rate. The regulatory influence of Cic is incorporated in a special way. Cic has a uniform dorso-ventral expression profile at the appropriate developmental stage, but a spatially patterned post-transcriptional modification of this protein is hypothesized to shape its regulatory effect on *ind*. In particular, a recent study (Lim, Samper et al. 2013) suggests that locally activated ERK reduces DNA binding by Cic, and that the resulting gradation in Cic-DNA binding may set the dorsal boundary of *ind.* The strength of this reduction is unknown and is one of the 13 free parameters of our model. We modeled the *ind* enhancer sequence along with the *ind* expression pattern in wild type as well as under three different perturbations (involving the three repressors Sna, Vnd, Cic) to determine the extent to which the model parameters can be constrained by the rich data available on regulation of this gene.

Our results can be summarized as follows. We found that a number of distinct quantitative models are consistent with the wild type *ind* pattern and that additionally modeling the gene's response to genetic perturbations of repressors substantially reduces the number and diversity of plausible models. Among other things, these models support and quantify the hypothesized interplay between ERK and Cic in defining the dorsal boundary of *ind* expression. At the same time, the ensemble of models makes clear predictions about the quantitative roles of Dl and Zld, underscoring previous observations that tight predictions can emerge from systems biology models despite diversity in parameter assignments (Gutenkunst, Waterfall et al. 2007). We experimentally tested and confirmed one of these predictions, viz., a reduction in *ind* expression upon Zld site mutagenesis, demonstrating that sequence-to-expression models can be used to make quantitative predictions in perturbation conditions. The learned models also explain why a recent study (Garcia and Stathopoulos 2011) found no significant change in *ind* expression upon mutating a high-affinity Dl site in the *ind* CRM, even though prior evidence suggests Dl as a major activator of this gene. Finally, we noted examples of situations where predictions from a model ensemble exhibit large uncertainty, opening the door for principled approaches to design of experiments that may significantly reduce this uncertainty.

In summary, we use sequence-to-expression modeling to determine what can and cannot be inferred, from data, about the *cis*-regulatory logic of a gene that is regulated by the combinatorial interplay of a morphogen, a signaling pathway, and both graded and uniformly expressed regulators. In doing so, we also hope to establish an example of how a quantitative study should proceed in developing a truly predictive model of gene expression from its regulatory sequences.

3.2 Results

3.2.1 A model of transcriptional regulation by transcription factors and their interplay with signaling molecules

In this work we modified GEMSTAT (He, Samee et al. 2010), a previously reported sequence-to-expression model, to study how TFs bound to the *ind* CRM may regulate the gene's expression. We outline the model here, see Methods for the details. GEMSTAT is founded on a theory of combinatorial gene regulation first proposed by Shea and Ackers (Shea and Ackers 1985). The model considers the system of TF molecules and their cognate sites in the CRM, as well as the basal transcriptional machinery (BTM) and its binding to the promoter, and uses a minimal set of parameters to model the interactions among TFs, BTM and DNA (Fig. 3.1-A). The model includes two parameters for each TF, that quantify the TF's DNA binding strength and its interaction with the BTM, respectively. Additionally, there is one parameter for each pair of TFs that are assumed to bind the DNA cooperatively, and one parameter to represent the basal expression level. All interactions are assumed to happen in thermodynamic equilibrium, which is reached much more rapidly than the time scale at which the transcription machinery is activated and begins producing mRNA. Under these assumptions, the transcription initiation rate, and hence the equilibrium level of mRNA transcription, is proportional to the fractional occupancy of the BTM at the promoter. The GEMSTAT model computes this fractional occupancy by considering all possible configurations of DNA-bound TFs and BTM (Fig. 3.1-B) and summing the probability of configurations where the BTM is promoter-bound (Fig. 3.1-C). The equilibrium probability of each configuration is computed as per the Boltzmann distribution (see Methods). As TF concentrations change across cell types, the probability of bound BTM configurations also changes, reflecting the variation of readout levels due to the change in regulator concentration (Fig. 3.1-D).

An important distinction of the GEMSTAT model from several other thermodynamics-based models (Zinzen, Senger et al. 2006, Zinzen and Papatsenko 2007, Papatsenko and Levine 2008, Gertz, Siggia et al. 2009, Fakhouri, Ay et al. 2010, Parker, White et al. 2011, White, Parker et al. 2012) is its ability to automatically account for varying affinities of a TF's binding sites, by relating mutations from the optimal or 'consensus' site to corresponding changes in binding energy. For this, GEMSTAT implements Berg and von Hippel's theory of protein-DNA interaction energetics (Berg and von Hippel 1987), using the TF's position weight matrix (Stormo 2000) to predict the "mismatch energy" relative to the consensus site (see Methods). In this work, we further extended GEMSTAT to allow for modulation of a TF's DNA binding affinity depending on the concentration of some other molecular species, which in our case (next section) was an extracellular signal regulated kinase. We used the newly implemented mechanism to model a "de-repression" effect whereby the kinase attenuates a repressor TF's DNA-binding affinity, resulting in higher expression levels of the regulated gene at higher levels of the kinase (Fig. 3.1-E). This allows us to model, for the first time, how a patterned but non DNA-binding regulatory input may shape the expression pattern of a specific target gene by interacting with the gene's enhancer.

3.2.2 A model of transcriptional regulation of the *intermediate neuroblasts defective* gene

We used the GEMSTAT model introduced above to study the details of regulation of *intermediate neuroblasts defective* (*ind*), a dorso-ventral (D/V) patterning gene in *Drosophila*. Our main goal was to characterize the roles played by various previously reported regulators of this gene, at a qualitative as well as quantitative level, infer mechanistic details of the combinatorial action of these regulators, and test if these details are consistent with observations made under various perturbations (*cis* as well as *trans*) of the system. Expression patterns of *ind* and its regulators are shown in Fig. 3.2-A. We begin here by listing the qualitative features of our model (Fig. 3.2-B), based partly on evidence from the literature.

- Dl activates *ind* (Hong, Hendrix et al. 2008) while Sna and Vnd work as its repressors (McDonald, Holbrook et al. 1998, Weiss, Von Ohlen et al. 1998, Cowden and Levine 2003) (Fig. 3.2-B,C).
- Zld activates *ind* (Nien, Liang et al. 2011)(Fig. 3.2-B,C). Additionally, we modeled Zld and Dl as exhibiting cooperative DNA binding at closely located binding sites. We noted the presence of five Zld sites in the *ind* CRM, with two pairs of adjacent Dl-Zld sites located < 25 bps apart (Fig. 3.2D), and similarly spaced Dl-Zld sites in orthologous sequences in other *Drosophila* species (Supplementary Fig. 3.1), suggesting the inclusion of Dl-Zld cooperativity in our model. Dl-Zld cooperativity is expected to allow the uniformly expressed Zld to accentuate Dl activation and could in principle lead to a steeper dorsal boundary of *ind* expression (Kanodia, Liang et al. 2012), mirroring a similar mechanism in the *sog* CRM (Liberman and Stathopoulos 2009). Including Dl-Zld cooperativity can also act as a surrogate for chromatin-mediated effect of Zld on Dl activation, as suggested in our recent work (Cheng, Kazemian et al. 2013). On the other hand, modeling a direct activating influence of Zld would increase the predicted peak expression levels but also increase the basal levels commensurately. We included both of these mechanisms (direct activation as well as indirect activation through cooperative binding with Dl) as model features, whose quantitative importance was left to be learned from the data.
- Cic acts as a repressor of *ind* (Fig. 3.2-B,C). It has been noted that *ind* expands dorsally upon mutating Cic sites in its CRM (Lim, Samper et al. 2013). However, Cic has a spatially uniform nuclear concentration during the pre-gastrulation stage (Lim, Samper et al. 2013), suggesting that an additional input that localizes Cic's activity domain must be considered when modeling Cic-mediated repression of *ind*. EGFR signaling may provide this input, presumably by relieving Cic-driven repression of *ind* (Cornell and Ohlen 2000, Hong, Hendrix et al. 2008, Chopra and Levine 2009, Ajuria, Nieva et al. 2011, Lim, Samper et al. 2013). In particular, EGFR is known to activate ERK (Ajuria, Nieva et al. 2011), which phosphorylates Cic and has been proposed to influence Cic activity by impeding its DNA binding (Dissanayake, Toth et al. 2011, Lim, Samper et al. 2013), leading to *ind* de-repression in a specific domain along the D/V axis. This is the mechanism we chose to implement here, though other mechanisms have also been proposed (Grimm, Sanchez Zini et al. 2012). We obtained a D/V profile of dual-phosphorylated ERK (dpERK) from (Lim, Samper et al. 2013) to serve as a surrogate for ERK activity. To model the interplay of EGFR signaling and Cic-driven repression, we modified GEMSTAT so that the energy of Cic-DNA binding is increased (binding affinity is reduced) to an extent proportional to dpERK concentration (see Methods for details).

We instantiated the GEMSTAT model using the above features. This model has 13 free parameters: two per TF representing its DNA-binding and activation/repression potency (denoted by K and α, respectively), one for Dl-Zld cooperativity (denoted by ω), one representing basal transcriptional activity (denoted by q_{BTM}), and one representing the attenuation of Cic's DNA-binding energy in proportion to the nuclear concentration of dpERK (denoted by Cic_{ATT}). Acquisition of data on the mRNA expression profile of *ind* and the concentration profiles of the regulatory TFs and of dpERK are described in Methods. The free parameters of the model were trained on the wild-type D/V expression profile of *ind*, and prediction from the trained model was found to be in excellent agreement with this wild-type expression pattern and also to be sensitive with respect to most of the parameters (Fig. 3.2-E), indicating that the model is flexible enough to capture the combinatorial effect of the assumed regulators in driving *ind* expression. However, this also raised questions about the validity and utility of the trained model, such as: 1) *Can the model correctly predict the effect of cis- and trans- perturbations to the system?* 2) *Is there a unique set of parameter values determined by the data?* 3) *What insights do the trained parameter values provide about the underlying mechanisms of* ind *regulation?*

3.2.3 Systematic exploration of parameter space provides an ensemble of plausible models that explain wild-type data

In Fig. 3.2-E, we presented the prediction of a single model, i.e., one particular setting of parameter values, that accurately fits wild-type *ind* expression. Any assignment of values to the 13 free parameters of the model corresponds to a predicted readout of the *ind* CRM, which can then be scored against the wild-type *ind* pattern using an appropriate "goodness-of-fit" function (see Methods). A high-scoring parameter assignment represents a plausible quantitative model of *ind* regulation, and its examination may provide insights into the relative strengths of various regulatory functions that are combined in the model.

Given any initialization of parameter values, the GEMSTAT program systematically and iteratively modifies those values and reports a locally optimal parameter setting that maximizes the goodness-of-fit. However, there may exist many other parameter assignments that are as good or nearly as good in terms of their agreement with data, and examining the one optimal assignment reported by GEMSTAT may provide a skewed view of plausible models (Kirk, Thorne et al. 2013). We therefore modified the GEMSTAT program to perform a comprehensive exploration of the multi-dimensional parameter space, with the goal of constructing a complete map of plausible quantitative models. To this end, we first generated a large number of 13-dimensional vectors (parameter assignments) as follows (Fig. 3.3-A). We partitioned each parameter's allowed range into two halves, which gave us 2^{13} compartments of the parameter space (see Methods). From each of these compartments, we sampled and scored 1000 vectors of parameter values for their goodness-of-fit to data. We next sorted the 1000 x 2^{13}, i.e., ~8 million, sampled parameter vectors based on their scores. Finally, for each parameter vector whose score ranks among the top 2% of unique scores in this sorted list (~21000 in total) we optimized the GEMSTAT model using that vector as initial estimate of model parameters. (See Methods for details.) The collection of optimized models can predict *ind* expression accurately in wild-type condition, with little dispersion in their predictions. We call this collection of models the "wild type ensemble". Interestingly, the ~21,000 models spanned widely different compartments of the parameter space (652 out of $2^{13} \approx 8000$). This suggested there might be many

distinct parameterizations that explain the wild-type data equally well. This is also apparent from the high variance and multimodality of the marginal probability densities of model parameters as estimated from the wild-type ensemble (Fig. 3.3-C, dotted curves). This however raises the concern that not all of these ~21,000 models would be able to yield accurate predictions under conditions different from the wild-type. As described in the next subsection, upon checking model predictions under different *trans*- and *cis*-perturbations, we could discard nearly 90% of models in the wild-type ensemble.

3.2.4 Data from perturbation experiments narrow down the range of plausible models

Wild-type data may not have sufficient information to constrain a highly flexible (parameter-rich) model into capturing the precise extent of each TF's effect on the target gene. To further constrain the values of model parameters, we examined how well models in the ensemble predict the effects of the following genetic perturbations for which we have data from the literature.

- Mutation of *sna*: the *ind* expression domain remains essentially unaltered in *sna* mutants. *vnd* expression is de-repressed in these embryos and expands ventrally such that *ind* stays repressed in the endogenous domain of *sna* expression.
- Mutation of *vnd*: the *ind* expression domain expands ventrally, yet does not encroach into the mesoderm region, in *vnd* mutants (McDonald, Holbrook et al. 1998, Weiss, Von Ohlen et al. 1998).
- Mutation of Cic binding sites in the *ind* CRM: the readout of the *ind* CRM expands dorsally, to an extent that matches the spatial domain of the Dl protein, upon mutating two particular Cic sites in the CRM (Lim, Samper et al. 2013).

These are the only perturbation results that manifest direct effects on *ind* expression. (See Discussion.) We used our model to predict *ind* expression pattern upon knocking down a TF: first, the DNA-binding parameter of the respective TF was set to zero (to simulate the absence of the TF) and secondly, if the knock-down affects the expression of a second regulator of *ind*, then the affected expression pattern was replaced with its altered expression pattern in the model. The latter case occurs in *sna* mutants, where the spatial pattern of *vnd* and *egfr* are altered. To predict the effect of mutating a site, we discarded the site from our set of annotated TF binding sites in the *ind* CRM (see Methods). In evaluating trained models on perturbation data we focused on carefully selected domains along the D/V axis which, we reasoned, should provide adequate information about the accuracy of model predictions (see Methods).

For each of the ~21,000 models in the wild-type ensemble, we evaluated its predictions on perturbation data and discarded every model that failed to correctly predicted the known effects. We found ~2100 models whose predictions are accurate in both wild-type and in the three perturbation conditions (Fig. 3.3-B). We call these models the "final ensemble". Parameters of these models were found to be far more constrained than those of the initial ensemble (Fig. 3.3-C, solid curves compared to dotted curves). One common class of models discarded in this step was those estimating a very weak activating input from Dl (low K_{DL}, α_{DL}) to *ind*. In fact, this class of models overestimate the activating role of Zld on *ind.* Therefore, these models consistently and incorrectly predict a high expression level of *ind* in the dorsal-ectoderm when Cic sites are mutated, leading to their exclusion from the final ensemble and suggesting that the filtered models have a more delicate balance between the activator and repressor parameter values than one may achieve solely by fitting wild-type data.

The additional constraints imposed above clearly narrowed our estimates for the ranges of several parameter values, but did not provide a unique model of *ind* regulation. In particular, model parameters in the final ensemble are located in 42 (out of 2^{13}) different compartments of the parameter space with large variability remaining in the activation-related parameters, namely K_{ZLD}, α_{DL}, α_{ZLD}, ω, q_{BTM}, and the repression-related parameter α_{SNA}.

3.2.5 Predicting the effect of mutating activator binding sites

We used the final ensemble models to investigate a major unanswered question about *ind* regulation: the relative contributions of its two activators – Dl and Zld. *ind* expression is known to become abolished in Dl mutants (von Ohlen and Doe 2000) and to become significantly weak in Zld mutants (Nien, Liang et al. 2011). However, both Dl and Zld are also implicated in regulating several direct regulators of *ind* (e.g., *sna, vnd,* and *egfr*) (Hong, Hendrix et al. 2008, Nien, Liang et al. 2011), so their genetic effects on *ind* may be a combination of direct and indirect influences. To accurately characterize the direct activating roles of Dl and Zld one needs to mutate their binding sites in the *ind* CRM. A computational scan of the *ind* CRM using PWMs of Dl and Zld identifies several putative binding sites for both TFs (Fig. 3.4-A; see Methods). Below we describe the final ensemble predictions of mutating these activator sites and relate the predictions to experimental data.

The only experimental study that examines direct effects of Dl on *ind* is that of Garcia and Stathopoulos (Garcia and Stathopoulos 2011), who mutated a Dl binding site in the *ind* CRM and found no significant change in *ind* expression. Predictions from our final ensemble for the particular mutation performed in (Garcia and Stathopoulos 2011) agree with that study, in that no effect on *ind* expression is predicted (Fig. 3.4-B). Importantly, we also predict that removal of all putative Dl sites will cause an abolishment of *ind* expression (Fig. 3.4-C). We consider this as an important quantitative explanation to Garcia and Stathopoulos' report, since it is impossible to predict purely based on qualitative reasoning that mutating the strongest Dl binding site in the CRM will have no significant effect on the gene's expression. However, our results also suggest Dl as a strong activator of *ind*, which Garcia and Stathopoulos could not affirm due to their focus on a specific binding site.

We also used the final ensemble to predict that Zld-induced activation is necessary for wild-type *ind* expression levels; specifically, that activation of *ind* should reduce to ~50% of its peak wild-type level upon mutating Zld binding sites in the CRM (Fig. 3.4-D). We tested this prediction experimentally, and noted that *ind* expression indeed reduces in transgenic embryos where Zld sites were mutated (Fig. 3.4-E). In particular, the mean intensity of *LacZ* mRNA expression upon mutating Zld sites dropped to an extent that agrees remarkably with our model prediction (Fig. 3.4-F). However, the expression of *ind* in these embryos appear to be noisier than its endogenous expression, leading us to speculate whether the apparent reduction in expression level is due to increased noise (i.e., *ind* is expressed either at a basal level or at a level comparable to its endogenous expression) or due to an overall reduction in expression level within the nuclei where *ind* is expressed. While the latter highlights a more direct transcriptional role of Zld, the former may be an artifact of Zld working a chromatin remodeler. Our analysis shows an overall reduction in the *LacZ* mRNA intensity (Fig. 3.4-G), suggesting a more direct involvement of Zld in activating *ind*. A possible link between this phenomenon and our model predictions is provided by the theoretical work of

Raser and O'Shea (Raser and O'Shea 2004). Raser and O'Shea proposed and validated a stochastic model (Raser and O'Shea 2004) that formulates the intrinsic noise in gene expression as a function of the rates at which a promoter switches between active and inactive states. For a gene where (a) promoter activation is infrequent relative to transcription and (b) the active promoter state is stable, their model predicts that a decrease in the rate of promoter activation should result in an increase in the intrinsic noise in gene expression. We note that, the aforementioned conditions are assumed to hold true in our model (more generally, in the Shea-Ackers' model). As noted above, our model predicts ~50% reduction in *ind* expression upon mutating Zld sites. More specifically, it predicts a lower BTM occupancy at the promoter when Zld sites are mutated, which in the parlance of Raser and O'Shea is a reduction in the rate of promoter activation and should result in increased noise. Thus, our model predictions combined with the Raser-O'Shea model (Raser and O'Shea 2004) suggest an explanation to the observed weak and sporadic *ind* expression resulting from Zld site mutations. Importantly, our prediction of weak *ind* expression complies also with the emerging view of Zld's function as a chromatin remodeler (Foo, Sun et al. 2014). In particular, one way to realize Zld's activating role in our model is as a facilitator of Dl's DNA binding (through the assumed cooperativity between Dl and Zld), a phenomenon reported recently for other D/V CRMs (Foo, Sun et al. 2014). To our knowledge, ours is the first model to confirm Zld's involvement and quantify its role in transcriptional *cis*-regulation.

3.3 Methods

Experimental methods specific to the results discussed in this chapter are given below.

3.4 Discussion

Recent advances in high-throughput and high-resolution assays have made quantitative biology rife with hypotheses that are often impossible to reconcile. Multi-parameter computational models with complex structures are currently the only means to unify these hypotheses into comprehensive descriptions of the biological systems under study. Parameter richness and structural complexity are likely indispensable aspects of these models since biological systems, as we understand them today, comprise many components working under a variety of interactions. Concomitant to these models is the issue of parameter uncertainty that makes it questionable to rely, for predictive purposes, on point estimates of the model parameters. The same has been demonstrated for computational models of transcriptional *cis*-regulation – the class of models we used in this study. In fact, an emerging perspective suggests parameter uncertainty is "universal" in systems biology models. Despite the concerns of parameter uncertainty, perhaps in response to them, *ensemble modeling* has been shown to be a powerful strategy for reducing prediction uncertainties in models of signal transduction networks (as well as those of climate change and protein folding). A major contribution of our work is the demonstration, for the first time, of how ensemble modeling may benefit a parameter-rich model of transcriptional *cis*-regulation, helping refine its parameter estimates. In doing so, we also show how such modeling can help us comprehend the disparate experimental evidence pertaining to regulation of the *ind* gene in *Drosophila*.

What if we did not adopt the ensemble approach? Standard approaches of finding one or a few optimal solutions do not guarantee finding *any* of our final ensemble models, regardless of whether one adopts a

global parameter estimation strategy or a local strategy coupled with random restarts. Given the sheer number of "wrong" models we discarded upon filtering with perturbation data, all of which were consistent with wild-type data, it is arguable that the conventional approaches are prone to report the wrong models and elicit incorrect inferences thereby. (The same was reported recently (Dresch, Liu et al. 2010) for a model of *cis*-regulation where random-restart based local parameter search repeatedly produced very similar parameter estimates (Zinzen, Senger et al. 2006).)

The benefits of ensemble modeling may not be immediately clear given that our final ensemble models make relatively "tight" predictions. We note that our final ensemble embodies a catalogue of many possible explanations to the details of *ind* regulation, although in the course of computing the ensemble we have discarded numerous explanations that did not meet additional consistency requirements imposed by perturbation data. By using the final ensemble models to make *in silico* predictions for novel perturbations, one may pose and experimentally validate (or reject) further hypotheses about the roles of different regulators of *ind*. For example, the final ensemble models predict a wide variation in *ind* expression in the dorsal ectoderm region upon mutating all Cic sites in the ind CRM. (Of note, the available experimental result of Cic site mutation was obtained by mutating two Cic sites in the CRM.) As such, the final ensemble suggests mutation of all Cic sites as an immediate experiment that can clarify the extent of Zld-mediated activation on *ind*.

An important finding in this work is the involvement of Zld in activating *ind*, with the broad message being that Zld binding is important for the establishment of wild-type *ind* expression. To our knowledge, this is the first demonstration, quantitatively under a combined modeling-experimental assessment, that Zld impacts transcription of a target gene. Current literature places Zld as a "potentiator" of activation by Dl, ostensibly by facilitating the DNA binding of Dl (Foo, Sun et al. 2014). Although cases are known where Zld binding sites do not follow a precise "grammar" and where the expression of Dl targets are delayed in *zld* mutants, a direct activating role of Zld has always been ruled out because none of the targets of Dl are expressed in genetic backgrounds lacking nuclear Dl. Our final ensemble models, however, suggest that both these roles of Zld may explain the available data. Thus, CRM-bound Zld may facilitate the binding of Dl, a proposition that supports Zld's role as a chromatin remodeler or as a functional associate of Dl, or it may directly activate *ind*, a proposition that suggests Zld has a role similar to other direct activators of gene transcription (e.g., Bicoid and Dorsal).

We also demonstrated here how weak binding sites may influence transcriptional activity. The role of Dl in activating *ind* was not clear either under the affinity-threshold model or based on the limited experimental results (Garcia and Stathopoulos 2011). Our analysis not only reconciled the previous experimental result with the existence of multiple weak Dl binding sites, but also makes a case about the functional importance of weak binding sites (Ramos and Barolo 2013).

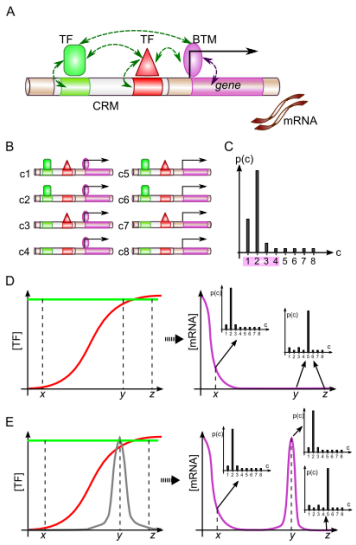
An important step in this study was the assumption that post-translational modification of CIC by ERK may inhibit DNA-binding by CIC and relieve *ind* from CIC-mediated repression in a manner that depends on ERK's spatial distribution. Without such an assumption, we were not able to fit any model with reasonable accuracy: models would estimate very weak repressive potential of CIC, but that also caused dorsal expansion of *ind*. To our knowledge this is the first demonstration of how signal transduction may

influence transcriptional regulation at the sequence level (Barolo and Posakony 2002). However, our goal was merely to capture a de-repression effect based on spatial distribution of ERK and our specific mechanistic assumption, although plausible based on several recent studies (Dissanayake, Toth et al. 2011, Lim, Samper et al. 2013), need not be the only way to achieve this de-repression. For example, alternative hypotheses about modifications in the influence of CIC on transcription initiation or on activator recruitment (without any modification in its DNA binding) are also plausible. Ascertaining any such mechanism is a subject for future studies.

Finally, it is worth discussing that although the Shea-Ackers' model assumes gene expression to be proportional to the total probability of BTM-bound configurations, this aspect of the model need not be interpreted literally. In particular, the BTM-bound configuration may be considered as a surrogate for a more complex biochemical state that is a pre-requisite for transcription initiation. For instance, recruitment of key co-activators or a critical chromatin remodeling event may be subsumed in the definition of the BTM-bound configuration, as long as the assumption of thermodynamic equilibrium can reasonably be made (Parker, White et al. 2011). Moreover, the Shea-Ackers model does not ignore events following BTM recruitment, which include isomerization of closed BTM-DNA complex to an open state and promoter clearance, among others. Rather, these events are modeled as one irreversible reaction with first order kinetics (Swain, Elowitz et al. 2002). The details of these events are not considered critical to model the effect of TFs on the overall transcription rate.

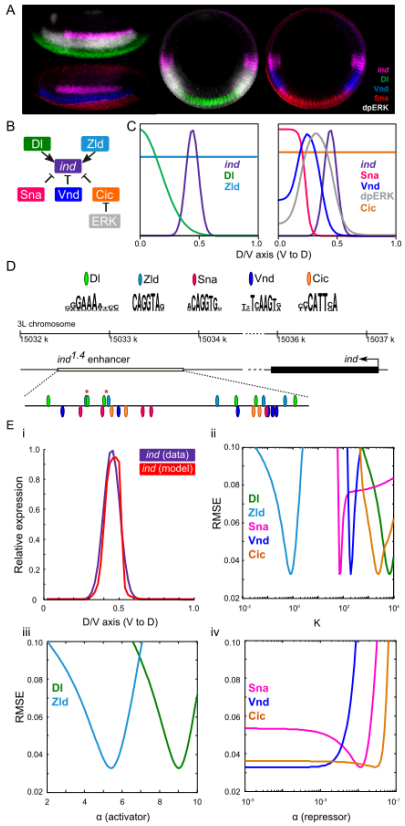
3.5 Figures

Figure 3.1: Overview of thermodynamic modeling of gene expression from enhancers



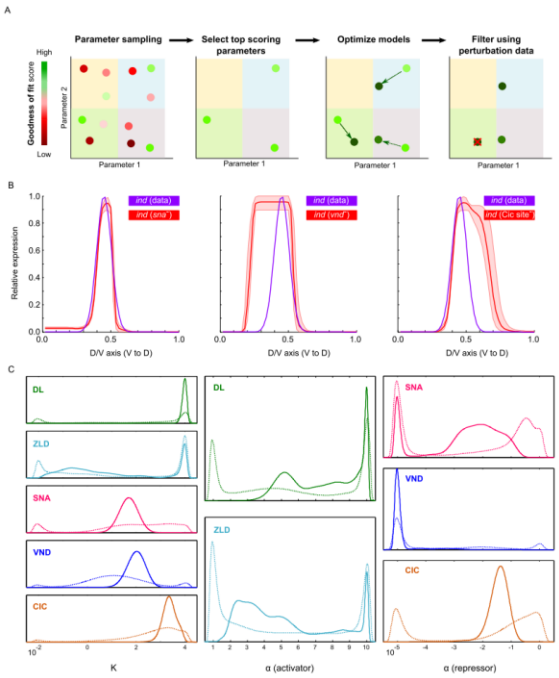
Overview of GEMSTAT. (A) GEMSTAT models the major components and their interactions involved in transcriptional regulation: the CRM (DNA sequence), transcription factors (TFs), and the basal transcriptional machinery (BTM). TFs bind at their cognate sites in the CRM and the BTM binds at the promoter. The mRNA expression level is determined by strength of TF-DNA interactions (at the binding sites) and TF-BTM interactions. Different possible interactions are shown with arrows. (B) GEMSTAT assumes that the system is at thermodynamic equilibrium. An exponential number of possible configurations of bound TFs and the BTM may occur in equilibrium. Shown are the eight possible configurations corresponding to the example shown in A. (C) GEMSTAT assumes the mRNA expression level is proportional to the equilibrium probability of the BTM binding at the promoter. Under standard statistical mechanical assumptions, equilibrium probabilities of configurations follow Boltzmann distribution. Shown is a hypothetical probability distribution for the configurations shown in B. The probability of BTM binding at promoter is computed from the probabilities of all BTM-bound configurations (i.e., configurations c1—c4). (D) GEMSTAT's predictions for mRNA levels change as the TF concentrations change across different experimental conditions. Shown is the profile of mRNA levels resulting from a uniformly expressed activator (green) and a graded repressor (red). The horizontal axis represents different conditions, e.g., different spatial locations along Drosophila D/V axis. Also shown are the different equilibrium probability distributions as the TF concentrations change at different conditions x, y, and z. (E) A molecular species (gray) that can attenuate the DNA-binding affinity of a repressor may also affect the mRNA level of the gene shown in D. Shown is how one such molecular species may de-repress the gene and result in high mRNA levels in conditions where the gene was previously repressed. Also shown is how the equilibrium probability distributions in D are affected due to this new molecular species.

Figure 3.2: Wild-type data and results of fitting GEMSTAT on wild-type data.



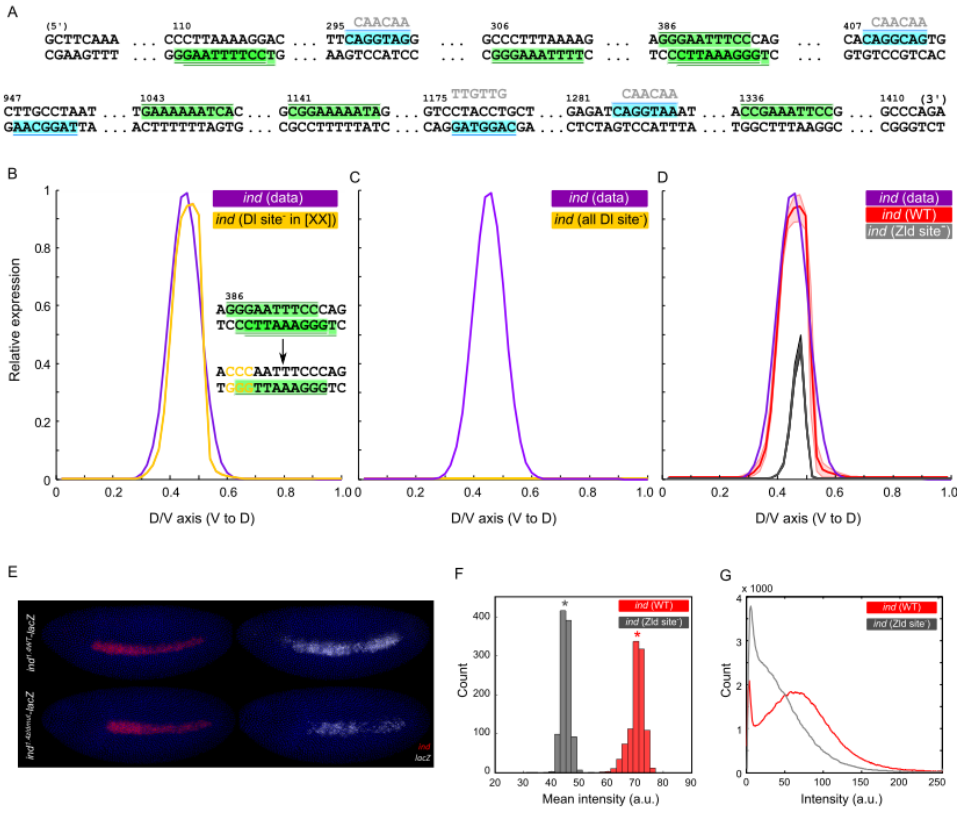
(A) Lateral (left) and D/V cross-sections (right) of *Drosophila* embryos in blastoderm stage. Embryos were stained with *ind* mRNA (magenta) and its four non-uniform regulators, Dl (green), Vnd (blue), Sna (red), and dpERK (gray). Two uniform regulators Zld and Cic are not shown. (B) The assumed regulatory relationships between *ind* and its regulators. Dl and Zld activates *ind*, while Sna, Vnd, and Cic are repressors. ERK represses Cic and thus relieves *ind* from repression. (C) Quantitative expression profiles of *ind* and its regulators along the D/V axis (computed using fluorescence data from images of > 10 embryos for each TF; see Methods). (D) PWMs used to scan the *ind* CRM and the map of TF binding sites matching the PWM motifs. Asterisks mark the two pairs of closely located Dl-Zld sites. (E) (i) Predicted *ind* expression from a model optimized on wild-type data. Purple and red curves show the wild-type data and the model prediction, respectively. (ii – iv) Sensitivity of a model (sampled from the wild-type ensemble) with respect to different parameters. Each panel shows the RMSE score of the model (vertical axis) as the corresponding parameter's value is varied from the minimum to the maximum of its range (horizontal axis), keeping other parameters fixed at their optimized values. For brevity, we limit the vertical axis at RMSE = 0.10.

Figure 3.3: Outline of ensemble construction, predictions from final ensemble models, and parameter values before and after filtering.



(A) Schematic of parameter space exploration and ensemble construction for a two-parameter model. Each parameter axis is partitioned into halves. A large number of samples are drawn from each of the four resulting compartments in the parameter space. Samples are scored according to their goodness of fit with respect to the data and the best samples are retained for further optimization in the second phase. In the second phase, we optimize one model, starting from each parameter retained in the first phase and the resulting optimized models constitute our wild-type ensemble. Models in the wild-type ensemble are filtered according to their accuracy in predicting the effects of various *cis*- and *trans*-perturbations. The remaining models (not crossed-out) constitute the final ensemble. (B) Predictions of the final-ensemble models under perturbed conditions: *sna* mutants (left), *vnd* mutants (middle), and mutation of two Cic sites of the *ind* CRM (right). Shown is the mean expression (red) and the standard error (shaded red area around the curve) for 1000 models sampled from the ensemble (we first sampled a compartment and then a model from the sampled compartment, both times uniformly at random). (C) Marginal densities of parameters of the wild-type and the final ensemble models (dashed and solid lines, respectively).

Figure 3.4: Predictions of the final ensemble models, and corresponding experimental results, upon mutating Dl and Zld sites in the ind CRM.



Semantics of the plots are the same as that in Figure 3B. (A) A computational scan of the ind CRM finds nine sites matching the Dl PWM (with two clusters of overlapping sites) and five sites matching the Zld PWM. Green and cyan boxes mark locations of Dl and Zld sites, respectively. (B) Final ensemble models do not predict any significant change in ind expression when the mutations performed by Garcia and Stathopoulos (Garcia and Stathopoulos 2011) are introduced to the CRM. Shown are Garcia and Stathopoulos' mutations and also the fact that although these mutations eliminate the targeted Dl site and two other sites overlapping with the targeted one, they create a new site for Dl. (C) Final ensemble models predict that ind expression abolishes upon removing all putative Dl sites. (D) Final ensemble models predict that ind expression reduces to ~50% of the peak expression upon mutating Zld sites in the ind CRM. The mutations are shown as gray sequences in A. (E) The ind CRM was used to drive expression that recapitulates the endogenous ind expression ($ind^{1.4WT}$-lacZ). Zld sites located in the CRM were mutated to study the Zld-dependent control of ind ($ind^{1.4zldmut}$-lacZ). Embryos were co-stained with ind (red) and lacZ (white). (F) Histograms of mean intensity values computed from 1000 bootstrapped intensity profiles of each type (i.e., wild-type and mutant; see Methods). Asterisks mark the bootstrapped mean values. (G) Smoothed histograms from the wild-type and mutant lacZ intensity profiles (see Methods), where each histogram was created from 20 profiles (one profile from each embryo) on 256 bins (one bin for each intensity value).

Chapter 4
Quantitative Modeling of a Gene's Expression from Its Intergenic Locus

4.1 Introduction

Gene regulation is key to understanding of a variety of biological processes ranging from development (Davidson 2006) to disease (Epstein 2009). Transcriptional regulation is one of the best studied stages of gene regulation (Courey 2008), especially in the context of developmental biology (White 2001). Studies of early embryonic development in *Drosophila* (Schroeder, Pearce et al. 2004) have revealed the roles of various transcription factors (TFs) in setting up precise spatio-temporal gene expression patterns, and delineated many "enhancers" (also called "cis-regulatory modules" or "CRMs") that mediate the activities of combinations of TFs. We have today a fairly detailed knowledge of the transcriptional regulatory network involved in patterning of the anterior-posterior (A/P) and dorso-ventral (D/V) axes in the blastoderm-stage *Drosophila* embryo (DePamphilis 2002, Arnosti 2003, Reeves and Stathopoulos 2009). This knowledge has spurred the development of quantitative models of gene regulation that aim to map the sequence of a given enhancer to the expression pattern driven by that enhancer (Buchler, Gerland et al. 2003, Bintu, Buchler et al. 2005, Janssens, Hou et al. 2006, Zinzen, Senger et al. 2006, Segal, Raveh-Sadka et al. 2008, Gertz, Siggia et al. 2009, Zinzen, Girardot et al. 2009, He, Samee et al. 2010, Kazemian, Blatti et al. 2010). These models attempt to (1) predict the strength of TF binding to sites within the enhancer by using data on TF concentration and binding specificity, and (2) integrate the predicted binding strengths of multiple TFs into a quantitative prediction of that enhancer's contribution to gene expression. The prediction may vary from one cell type to another, as TF concentrations vary. The ultimate goal is to build a computational tool that automatically predicts the expression of any gene in any cellular condition based solely on the genome sequence and a quantitative description of the trans-regulatory context (Kim, Martinez et al. 2013). Such a computational tool will embody our knowledge of the so-called "cis-regulatory code" (Istrail and Davidson 2005, Ochoa-Espinosa and Small 2006). It will help us annotate the regulatory genome at a single nucleotide resolution, and predict the effects of genotypic changes (in cis or in trans) on gene expression and phenotype.

Gene expression modeling: In this study, we consider the problem of *modeling* gene expression, which is an important intermediate step in the more ambitious goal of building the *predictive* tool mentioned above. In the modeling task, we are given the inputs (sequence and TF concentrations) and output (gene expression), and a model with tunable parameters is trained to map the inputs to the outputs. Such a model has many possible uses. Once trained on wild-type data on a gene, it can be used to predict outputs on non-wild-type inputs, which may include changes in cis (sequence) or trans (TF concentrations). It can provide a quantitative description of how a specific gene's regulation is encoded in the sequence, and can precisely characterize each TF's role in regulating the gene. Moreover, by testing alternative models in terms of their goodness-of-fit on the data, we can gain valuable insights into the mechanisms underlying gene regulation. Here, we develop such a model of gene expression, show that it fits the available data accurately, and demonstrate a few practical utilities of the model.

4.1.1 Locus-level gene expression modeling

A key challenge in achieving the above-stated goal is to model a gene's expression *from the sequence of its entire intergenic region,* or "locus". While regulatory influences on a gene have been known to be located at great distances (> 1 Mbp) from the gene (Sagai, Hosoya et al. 2005, Visel, Rubin et al. 2009), it is frequently observed that much of the information about the gene's expression pattern is encoded in its locus (Blackwood and Kadonaga 1998). This information is typically organized in modular units of length ~1 Kbp, called enhancers, that are scattered in the locus, both proximal and distal to the gene, and upstream, downstream as well as within introns of the gene. For instance, complex gene expression patterns such as the seven-striped patterns of "pair-rule" genes (Fig. 4.1-A,B) in the *Drosophila* embryo are known to be determined by multiple, distinct enhancers (Fig. 4.1-C), each of which is sufficient to drive a discrete aspect (one or two stripes) of the gene's overall pattern (Riddihough and Ish-Horowicz 1991, Andrioli, Vasisht et al. 2002). How the information encoded by multiple enhancers in a locus is integrated together is a largely unexplored problem. A simple hypothesis might be that the binding sites located across different enhancers in a gene's locus constitute one large enhancer, interpreted by the same rules of combinatorial action that apply to binding sites within any single enhancer. The more common view (Howard and Struhl 1990, Fujioka, Emi-Sarker et al. 1999), however, is that each enhancer is interpreted independently of others, and readouts of multiple enhancers are superimposed or combined additively to produce the gene expression pattern. If this latter view is more accurate, existing sequence-to-expression models, which have been tested on individual enhancers, may not suffice to model a gene's expression from its entire intergenic region. Indeed, while there have been several successful attempts to model enhancer readouts, especially for A/P and D/V patterning genes in *Drosophila* (Buchler, Gerland et al. 2003, Bintu, Buchler et al. 2005, Janssens, Hou et al. 2006, Zinzen, Senger et al. 2006, Segal, Raveh-Sadka et al. 2008, Gertz, Siggia et al. 2009, Zinzen, Girardot et al. 2009, Fakhouri, Ay et al. 2010, He, Samee et al. 2010, Kazemian, Blatti et al. 2010, Kim, Martinez et al. 2013), we are not aware of any computational model that has been successfully tested on a multi-enhancer gene locus such as those of the pair-rule genes (Figure 4.1). Our primary objective in this work is to implement and test such a computational model. A recent study by Kim et al. (Kim, Martinez et al. 2013) makes significant contributions to this modeling question, although the authors' primary focus was on elucidating specific details of transcriptional control mechanisms. (Also see Discussion.)

We present a computational framework for modeling the expression level of a gene from the sequence of its locus and a quantitative description of the trans-regulatory context (TF concentrations). We refer to this task as "locus-level gene expression modeling", where a gene's locus is considered to be the non-coding sequences extending upstream and downstream of the gene until a neighboring gene's boundary. (This includes UTRs and introns.) Our new model, called GEMSTAT-GL ("**G**ene **E**xpression **M**odeling based on **Sta**tistical **T**hermodynamics - **G**ene-locus **L**evel"), implements the two-layered, modular organization of cis-regulatory information mentioned above, thus reflecting the commonly held view today.

4.1.2 Practical problems in implementing a locus-level model

An important challenge for a model that interprets multiple enhancers in a locus and combines their separate readouts is the unknown location of enhancers in a locus (Halfon, Grad et al. 2002, Lifanov,

Makeev et al. 2003, Kazemian, Zhu et al. 2011) – an enhancer is typically ~1 Kbp long and may be located anywhere within the much longer (often 10-50 Kbp long) gene locus. Accurate identification of all the necessary enhancers in the locus will be a prerequisite for modeling gene expression. High throughput characterization of chromatin marks (Visel, Blow et al. 2009, Ernst, Kheradpour et al. 2011, Kharchenko, Alekseyenko et al. 2011, Negre, Brown et al. 2011) and computational enhancer scans (Berman, Nibu et al. 2002, Halfon, Grad et al. 2002, Frith, Li et al. 2003, Sinha, van Nimwegen et al. 2003, Donaldson, Chapman et al. 2005, Philippakis, He et al. 2005) may help overcome this challenge in the future; but ideally the quantitative model should automatically discover the contributing segments in the locus, rather than relying on enhancers identified *a priori*. A second major challenge in locus-level modeling is to model the mechanisms that integrate outputs from distinct enhancers into the endogenous gene expression. As noted above, a relatively simple 'additive' mechanism has been suggested in the literature (Howard, Ingham et al. 1988, Howard and Struhl 1990, Ishihara, Sato et al. 2008, Perry, Boettiger et al. 2011), where readouts of the contributing enhancers are summed up to produce the gene expression pattern. However, existing quantitative models often are capable of predicting enhancer readouts only on a relative scale (e.g., expression pattern along the A/P axis rather than absolute expression values). As such, it is not clear if a simple summation of model predictions on enhancers will suffice to accurately predict gene expression patterns. Moreover, while a minimal set of enhancers may capture all aspects of the gene expression pattern, it is not clear what role the rest of the locus plays. If the locus harbors multiple enhancers with similar readouts, as has been suggested by the discovery of "shadow enhancers" (Hong, Hendrix et al. 2008, Perry, Boettiger et al. 2010, Barolo 2012), a quantitative model should take into account contributions from all of them. These are some of the challenges related to locus-level gene expression modeling that motivate our work.

4.1.3 Overview of model development and testing

We report here the first general-purpose quantitative model of a gene's expression pattern as a function of the sequence of its entire locus. Here, "general-purpose" implies that the model can be trained on any given dataset with minimal or no manual parameter tuning. Admittedly, the model has to be provided with a complete set of candidate regulators (TFs), as well as their DNA binding motifs and relative concentrations, which currently limits its applicability to regulatory networks where such information is available. But given this information the model then automatically learns values for all of its free parameters, and the locations of relevant enhancers in the gene locus. As noted above, the new model treats the expression readout of an entire gene locus as being two tiered – 1) sites within each enhancer act together to produce that enhancer's contribution, which is modeled using the thermodynamics-based GEMSTAT model of enhancer function (He, Samee et al. 2010), and 2) contributions from multiple enhancers are combined as a weighted sum to produce the gene expression profile. We initially focused on the expression patterns of the genes *even-skipped, hairy, runt,* and *giant* in the developmental stage following the maternal to zygotic transition (DePamphilis 2002) in early *Drosophila melanogaster* embryos. In this stage, each of these genes is expressed in a complex multi stripe pattern and is known to be regulated by multiple enhancers within its locus, and is thus an ideal test case for locus-level modeling. As a point of contrast to the two-tiered model of GEMSTAT-GL, we also trained the GEMSTAT model that was shown previously to accurately model ~40 enhancers involved in A/P patterning. We found that

GEMSTAT fails to model multiple sharply defined stripes and instead predicts one broad expression domain when it is used for locus-level modeling of each of the four genes mentioned above. In order to demonstrate the broader applicability of GEMSTAT-GL, we next used it to model the expression patterns of 23 additional genes in early *Drosophila* embryo (we have thus modeled all the 27 A/P genes from Kazemian et al. (Kazemian, Blatti et al. 2010)). From the intergenic locus of each gene, our model automatically selected one or a handful of segments that together generated the gene's expression. The selected segments were found to overlap enhancers known to regulate the gene, even though the model was not informed about these enhancers. We also investigated whether and how the intergenic sequence outside these selected segments contributes to the gene's expression. Our findings suggest the presence of sequence segments in the locus that would exert an irreconcilable impact on the gene's expression pattern and thus were required to be explicitly "shut down" by the model, presumably reflecting a similar phenomenon *in vivo.*

4.1.4 Practical utilities of the new model

We used our models to analyze several aspects of the regulation of *eve*, *h*, *run*, and *gt*. 1) An immediate practical benefit of our model is the automatic discovery of candidate enhancers in the locus, along with accurate assignments of regulatory activity to each enhancer. This goes one step beyond our previous work (Kazemian, Blatti et al. 2010) where enhancers were annotated based on their pattern generating potential. The new method ensures that activities of multiple enhancers in the locus can be aggregated to match the gene's expression profile. Also, since GEMSTAT-GL allows model parameters to be trained simultaneously with the discovery of enhancers in a gene's locus, the assignment of regulatory activity to enhancers is empirically more accurate than those reported in (Kazemian, Blatti et al. 2010). 2) We performed *in silico* knock-downs of TFs and identified the TFs responsible for the formation of stripe boundaries in A/P expression patterns of these genes. The resulting network of regulatory interactions exhibits a very high level of agreement with known regulatory influences on the target genes, illustrating the potential of the model-based approach for unraveling regulatory networks. 3) We also developed a method to investigate whether and why the assumed independence of enhancers was necessary in our model. We found that interaction or 'cross-talk' (Kirstein, Sanz et al. 1996, Yao, Phin et al. 2008, Prazak, Fujioka et al. 2010, Perry, Boettiger et al. 2011) between the enhancers of a gene is detrimental to our model's fits to the gene's expression data, and identified cases where specific binding sites in one enhancer that may interfere with another enhancer's readout. This suggests that in these cases the independence of enhancer contributions is necessary for proper modeling of gene expression.

4.2 RESULTS

4.2.1 A thermodynamics-based model accurately predicts readouts of the enhancers of *even-skipped*, *hairy*, *runt*, and *giant*

We previously reported a statistical thermodynamics-based model of enhancer function, called "GEMSTAT", that was shown to successfully predict the expression patterns of ~40 enhancers of the anterior posterior (A/P) axis patterning system in early *Drosophila* embryo (He, Samee et al. 2010). GEMSTAT is built on basic physical principles laid out by Shea and Ackers (Shea and Ackers 1985, Buchler,

Gerland et al. 2003). It is the only available general purpose tool that can predict the expression readout of an arbitrary DNA segment and whose parameters can be trained on any given set of enhancers. It assumes gene expression in a cell type to be proportional to the fractional occupancy (Ay and Arnosti 2011) of the basal transcriptional machinery at the gene promoter, and estimates this occupancy from the enhancer sequence and the binding specificities (motifs) and concentrations of TFs in that cell type. Due to its previous successful application to individual enhancers and due to our extensive experience with it, GEMSTAT was a natural initial choice for modeling a gene locus. We made a major modification to GEMSTAT's objective function, which is used to compare predicted and real expression patterns: instead of the "root mean square error" function (He, Samee et al. 2010), it now uses a "weighted Pattern Generating Potential" (w-PGP) function (Samee and Sinha 2013) that was designed specifically for comparing spatial gene expression patterns. (See Chapter 2.)

Before using GEMSTAT to model the entire locus, we sought to confirm if it accurately models the characterized enhancers of the genes of interest in this study. We first focused on four genes in the early *Drosophila* embryo, namely *even-skipped (eve)*, *hairy (h)*, *runt (run)*, and *giant (gt)*. The multi-stripe patterns of these genes (e.g., Fig. 4.1-A,B) are among the first manifestations of complex combinatorial regulation in the *Drosophila* embryo (Reinitz and Sharp 1995). These genes are initially regulated by an interplay of maternally deposited proteins and their immediate regulatory targets (DePamphilis 2002), and their expression is later stabilized through more complex mechanisms including auto-(Harding, Hoey et al. 1989, Jiang, Hoey et al. 1991) and cross-regulation (Harding, Rushlow et al. 1986). Due to the complexity and multi-enhancer origins of their expression patterns and due to availability of high resolution expression data (Pisarev, Poustelnikova et al. 2009), these four genes were chosen as the primary subject of our study. A total of 18 early functioning enhancers have been reported in the literature for *eve*, *h*, *run*, and *gt* – 5 for *eve* (Fig. 4.1-C), 7 for *h*, 3 for *run*, and 3 for *gt* – each responsible for some discrete aspect (typically one or two "stripes") of the respective gene's pattern during early stages of development. Thus, for each gene our dataset included the sequences and known expression readouts of each enhancer, and the DNA motifs and A/P concentration profiles of nine TFs – BCD, CAD, ZLD, GT, HB, KNI, KR, TLL, and SLP (Fig. 4.1-D) – that are known to regulate expression at this stage of development (DePamphilis 2002, Harrison, Li et al. 2011, Nien, Liang et al. 2011). For each gene, GEMSTAT learns one set of parameters so as to maximize the agreement between predicted and known expression profiles of all enhancers of a gene according to the w-PGP metric (see Methods). As shown in Fig. 4.1-E, readouts of known enhancers were modeled accurately for each of the four genes, suggesting that the GEMSTAT model captures the combinatorial action of multiple, heterotypic binding sites in those enhancers. (The enhancers responsible for stripes 2, 4, and 6 of *run* are not known.) This exercise is shown schematically in Fig. 4.2-A. We used a constrained parameter estimation strategy here to guard against over-fitting. (See Methods.)

4.2.2 Intergenic locus readout under the thermodynamic model does not agree with multi-stripe expression pattern

Having confirmed that GEMSTAT can model enhancer readouts accurately, we next tested if GEMSTAT can model the multi-stripe patterns of the genes of interest from their respective intergenic regions (Fig.

4.2-B). By doing so, we hoped to answer the following question raised in the introductory section: Do the rules for interpreting a collection of binding sites in an enhancer apply unchanged to the larger collection of sites present throughout the locus? The intergenic region or "locus" was defined here as the sequence bounded by the immediate neighboring genes on either side (Fig. 4.1-C), and was of length 17 Kbp, 68 Kbp, 58 Kbp, and 17 Kbp for *eve*, *h*, *run*, and *gt*, respectively.

We performed two exercises, under different assumptions about the range of regulatory influence of repressors. In the first exercise, we assumed that repressor molecules bound to their cognate binding sites can directly affect the transcriptional machinery ("DIRECT INTERACTION" mode of GEMSTAT (He, Samee et al. 2010)). However, GEMSTAT was unable to find any set of parameters for which the predicted gene expression profiles match the multi-stripe profiles. One possible explanation for this failure is the phenomenon of "short range repression" (SRR). Some of the repressors of this regulatory system (e.g., GT, KNI, and KR) are known to act over short ranges only, i.e., their binding sites mediate a repressive action only if located within 100-150 bp of activator sites (Fakhouri, Ay et al. 2010). Therefore, in our second set of tests we trained GEMSTAT in the "SRR" mode (He, Samee et al. 2010), which captures short range repression, on each gene's locus (Fig. 4.2-C). However, this test was also unsuccessful, i.e., no parameter setting was found for which predicted expression profiles match the real gene expression profiles. We note that all of these failed experiments were performed with an unconstrained parameter estimation strategy (which is GEMSTAT's default strategy, see Methods). Therefore, failures of these experiments were presumably not due to shortcomings of the parameter optimization algorithm.

The finding that GEMSTAT successfully models enhancer functions but fails on the entire locus has at least two possible explanations. The first explanation is that binding sites within certain segments in the locus contribute to gene expression while sites outside of these segments do not contribute, and their inclusion in the model is somehow detrimental to the goodness-of-fit. To test this, we concatenated the known enhancers of each gene (Fig. 4.2-D) and searched for the best fit between GEMSTAT predictions and data. No satisfactory fit was found, suggesting that the above explanation is not sufficient. A second explanation for the failure of GEMSTAT on locus-level modeling has to do with the way GEMSTAT models the sequence. It computes the readout as a single non-linear function of (the strengths of) all binding sites in the sequence. Perhaps the readout of the locus is not best described as computing this function on all sites in the locus, even though the readout of individual enhancers does conform to this model. An emerging hypothesis was that local clusters of sites act together in ways captured by the GEMSTAT model (as demonstrated by the enhancer modeling exercise above) but contributions from different clusters of sites do not interfere with each other and these clusters should not be interpreted together. This hypothesis reflects the conventional wisdom about cis-regulatory architecture, and was reached here on the basis of the failed modeling exercises described above. We explored this hypothesis next, within a modeling framework, and found it being supported by all the genes modeled in this work.

4.2.3 A two-tiered model based on GEMSTAT accurately predicts expression from the entire gene locus

Our working hypothesis now was that distinct segments in the gene locus are interpreted separately based on the collection of sites within each segment, and their individual readouts are then aggregated to

produce the overall pattern. Thus, it presents a "two-tiered" gene expression model. The main challenges in formulating and training such a model are: (i) determining the segments whose readouts are aggregated, and (ii) choosing an appropriate aggregator function. The quantitative model may not assume prior knowledge of enhancers in the locus since such a strategy is not generalizable to poorly characterized loci. Gene expression profiles should be modeled solely from the gene locus and TF data (concentrations and motifs).

Pursuing the above hypothesis, we implemented a two-tiered model that uses contributions from a number of sequence windows in the locus, and predicts gene expression as a weighted sum of these contributions (Fig. 4.2 E). We call this new model "GEMSTAT-GL", with "GL" abbreviating for "gene-locus level". The sequence windows were allowed to be of varying lengths, even mutually overlapping if necessary, and their separate readouts were predicted using GEMSTAT. The number and locations of contributing sequence windows, as well as the weight of each window's contribution were left to be automatically discovered during model training. Model training was performed iteratively, with a new sequence window being included for contributing to gene expression only if its inclusion significantly improved the agreement between predicted and real expression profiles. In this way, the complexity of the model was kept under control. Details of this two tiered model and its parameter estimation procedure are described in Methods. Roughly speaking, this procedure (a) finds a window whose GEMSTAT readout matches one aspect (e.g., a stripe) of the gene expression pattern, (b) tests if a weighted summation of this window's readout and the readouts of already selected windows improves the overall prediction, and (c) includes the window if such an improvement is noted. The model parameters were fit separately for each gene; hence we adopted a "constrained" parameter estimation strategy to avoid over-fitting (see Methods and Discussion).

Predictions from the GEMSTAT-GL model agreed very well with the real expression profiles of each of the four target genes, *eve, h*, *run*, and *gt* (Fig. 4.3-A-D). For instance, we noted that the seven-stripes of *eve* and *h* expression were faithfully captured by the model (Fig. 4.3-A,B), while the seven-striped pattern of *run* was well approximated by a six-striped predicted pattern, with the model failing to separate stripes 4 and 5. Both domains of *gt* expression and their experimentally characterized assignments to three different enhancers were reproduced by the model. The agreement between model and data seen here for the *eve* and *h* stripes is qualitatively superior to corresponding fits in previous work on enhancer modeling (Segal, Raveh-Sadka et al. 2008, He, Samee et al. 2010), and this may be attributed to the fact that GEMSTAT-GL fits parameters on each gene separately. However, subsequent control experiments (described next) largely ruled out the possibility of obtaining such accurate models through over-fitting and highlighted the significance of the reported models. From each gene's locus, the model chose a small number of segments (at most seven) in the first tier before aggregating their GEMSTAT-based readouts in the second tier. The segments selected from a locus received comparable weights, with their values differing by at most two-fold. Moreover, these automatically chosen segments showed strong overlap with previously characterized enhancers of the respective genes (Fig. 4.3-A-D), even though the enhancers were not known to the model training procedure. In particular, of the 21 regulatory segments chosen from the four gene loci, 16 overlapped with REDFly enhancers (Gallo, Gerrard et al. 2011).

This initial success of the model motivated us further to test its generalizability. We therefore applied the model to all 27 A/P genes considered in (Kazemian, Blatti et al. 2010). These 27 genes, which are expressed between stages 4 and 6 during *Drosophila* embryogenesis, include several gap genes, pair-rule genes, and anterior, posterior, trunk, and terminal genes. They are, with the exception of secondary pair-rule genes (Text S1), likely to be regulated primarily by the maternal and the early zygotic proteins, and therefore are reasonable targets for modeling using the same input TFs as above. (We also used the TFs Capicua (CIC), Forkhead (FKH), and Huckebein (HKB) in modeling these genes, as in (Kazemian, Blatti et al. 2010).) The four genes modeled above – *eve, h, run,* and *gt* – are included in these 27 genes; hence we show the modeled expression patterns of the additional 23 genes in Fig. 4.4. GEMSTAT-GL was able to accurately fit the expression pattern for most of the genes, demonstrating its wide applicability for gene-locus modeling. The model fits were less accurate for the secondary pair-rule genes *ftz*, *odd*, and *prd*, where 4, 3 and 5 stripes were correctly reproduced (out of seven stripes of each gene). This relative lack of accuracy is probably because the direct regulators of these genes include the primary pair-rule proteins (Schroeder, Greer et al. 2011), which were not among the input TFs (see Text S1). Another case of model failure was *ttk*, presumably because the precise seven-striped pattern of *ttk* occurs later than stage 6 of embryogenesis (Brown and Wu 1993) and it requires other regulators than the used TFs (e.g., Biniou (Jakobsen, Braun et al. 2007)). To model these additional 23 genes, GEMSTAT-GL selected 29 regulatory segments, 23 of which were overlapping with REDFly enhancers. As above, a constrained parameter fitting strategy was used here.

4.2.4 Control experiments suggest that the trained model is not over-fit

Over-fitting was a concern in the above modeling exercise, since our framework does not allow testing of predictions on unseen data. We performed a number of control experiments, described next, to address this concern. As "negative controls", we repeated the above model-training exercise on the following types of artificial datasets (see Methods): (a) the locus of one gene was used to model the expression pattern of a different gene, (b) the locus of a given gene was used to model a "random" expression pattern, and (c) a gene's expression pattern was modeled from a randomly generated sequence of the same length as the gene's locus, and (d) a gene's expression pattern was modeled from a random relocation of TF binding sites in its locus. Each negative control experiment failed, as expected: no parameter settings were found for which model predictions agreed with data. Moreover, experiment (d) allowed us to assess the significance of our original model fits by comparing the goodness-of-fit score (value of objective function) of the trained model to an empirical distribution of scores from 100 negative controls for each gene. The original models were highly significant, with goodness-of-fit scores greater than all negative controls and with values 30-40 standard deviations above the mean from negative controls. We note that, as opposed to the constrained parameter estimation strategy in the modeling of real data, there was no constraint on parameter values in the control experiments. As an additional test, we trained the model on *D. melanogaster* gene expression profiles of *eve, h, run* and *gt* using sequence from the loci of their respective *D. pseudoobscura* orthologs. We assumed that the expression profile characterized experimentally in *D. melanogaster* remains unchanged in this related species (Segal, Raveh-Sadka et al. 2008). The trained model was found to capture the real expression profiles well (Fig. 4.3-E), although not as accurately as in *D. Melanogaster*: for the seven-striped patterns of *eve*, *h*, and *run*, the

model reproduced the locations of 6, 7, and 6 stripes respectively, though the inter-stripe boundaries were not as prominent as in the *D. melanogaster* models. The model fits on *gt* reproduced both anterior and posterior domains of endogenous expression, though the model-predicted domains were shifted posteriorly. We note again that we are unable to test the trained model by direct prediction of the readout of an unseen gene locus, since the locations and weights of contributing sequence windows have to be learned from that locus.

4.2.5 A sampling strategy reveals the cis-regulatory architecture of a gene locus

The two-tiered model described above discovered a small number of segments whose readouts could be aggregated to match the gene expression profile. This set of segments describes the "regulatory architecture" of the gene locus (Fig. 4.3-A-D), as a checkered pattern of putative enhancers (green boxes in the genome browser views) interspersed with large spacer regions that do not contribute to gene expression. However, since the model was trained with a local search algorithm and was designed to utilize only as many segments as necessary, it is possible that the learned architecture is one of many possible architectures, each of which has its own locations of putative enhancers and intervening spacers. To investigate this possibility, we performed Markov Chain Monte Carlo sampling of the space of architectures. (See Methods for details.) Each architecture was represented by the locations of sequence segments that contribute to gene expression, and their respective weights. Also, each architecture was sampled with probability proportional to its w-PGP score, which quantifies how well the model predictions for that architecture agree with gene expression. A summary of the large number (50,000) of architectures sampled by this scheme from the *eve* locus is shown in Fig. 4.5-A. It shows the average weight that a segment received over all samples. (A weight of zero indicates that the segment was part of the spacer regions between putative enhancers in that architecture, and weights cannot be negative.) We see that the average weights are heavily peaked at a handful of locations, while most other segments within the locus have very low average weights. Moreover, the high weight locations are coincident with the contributing segments from the optimal architecture found above (Fig. 4.3-A). This indicates the existence of a unique regulatory architecture at the gene locus. We also noted that the high weight segments of this architecture overlap known enhancers of the gene.

On the other hand, there were many segments with average weight close to 0 (Fig. 4.5-A), that were not included in any sampled architecture. Such segments either (a) have no regulatory information within them, or (b) their readout as predicted by the GEMSTAT model is inconsistent with and must not be aggregated with the readouts of other segments. The latter possibility suggests that there may be segments that exert an irreconcilable impact on the gene's expression pattern and thus have to be explicitly "shut down" by the model. A direct examination of their predicted readouts confirmed that this was indeed the case for some segments (Fig. 4.5-B). While most of the non-contributing segments had no noticeable readout, some such segments led to predicted expression at levels comparable to the known enhancers but at inappropriate axial positions, i.e., outside the stripe domains.

4.2.6 A regulatory network of transcription factors determining "stripes" of gene expression

One of the advantages of a quantitative model of gene expression is that it allows us to predict the effects of perturbations in *cis*- (the regulatory sequence) or in *trans*- (the transcription factors) on expression. For example, a "*knock-down"* of a TF is easily simulated by setting the TF's concentration to zero. Such *in silico* knock-downs may then be used to infer regulatory influences of any TF on the gene, and a transcriptional regulatory network may be constructed. In our past work (Kazemian, Blatti et al. 2010), we constructed such a regulatory network at the level of individual enhancers, i.e., the network predicted when a TF's knock-down would significantly affect an enhancer's readout. Such an effect does not necessarily translate to a change in gene expression, as there may be redundancy of information in the locus (Frankel, Davis et al. 2010, Perry, Boettiger et al. 2010, Barolo 2012). An advantage of having a quantitative model of the readout of the entire gene locus is that regulatory networks may be constructed at the level of genes rather than enhancers. An edge in such a network would correspond to a TF's knock-down affecting the gene expression; such an effect can be then be probed experimentally through an *in situ* hybridization assay in TF^- condition. (Testing a TF-enhancer association experimentally would involve reporter gene assays, which are more expensive.)

Here, we used *in silico* knock-downs to predict TF-gene regulatory interactions, and described the predicted interactions as a "TF-stripe" network where edges connect TFs to specific stripes in the gene's expression profile, reflecting an effect of the TF on establishment of that particular stripe. The TF-stripe network for the *eve* gene (Fig. 4.6-A) shows 35 edges (12 activating, 23 repressive influences) between nine TFs and seven stripes of *eve* expression. The activators BCD and CAD regulate the anterior and posterior stripes, as expected from their concentration profiles. Each of the two borders (anterior and posterior) of any stripe is regulated by one or two TFs. This automatically constructed network is in very high agreement with the literature: 30 of the 35 edges have been previously confirmed or hypothesized based on genetic evidence, and only two interactions (small dashed edges: BCD $\rightarrow$ Stripe 5 and HB $\rightarrow$ Stripe 2) with experimental evidence were not recovered by our procedure. The "HB $\rightarrow$ Stripe 2" interaction cannot be recovered by our model because we assign a fixed role (activator or repressor) to each TF, while the literature points to an activating role for HB at stripe 2 (Small, Blair et al. 1992, Arnosti, Barolo et al. 1996) and a repressive role elsewhere (Zhang and Bienz 1992). Overall, the strong agreement between the predicted and previously characterized TF-stripe network strongly argues for the usefulness of our approach, when we consider the vast amount of experimental work that has gone into characterizing those 30 recovered edges. Moreover, our model-based approach predicts three regulatory interactions that were not known previously (large dashed edges). These include roles for TLL and SLP in setting up the anterior border of Stripe 1 and a role for TLL at the posterior border of Stripe 5. Similar TF-stripe networks were constructed for *h*, *run,* and *gt*; these networks are shown, along with known interactions from the literature, in Fig. 4.6-B-D. As in the network for *eve*, we missed very few of the known edges in these latter three networks, and most of the missed edges correspond to 'indirect' activation (i.e., if A is a repressor of B and B is a repressor of C, then A indirectly activates C) which can only be captured by a network level model of gene regulation (Dresch, Thompson et al. 2013).

A comparison of the networks predicted by GEMSTAT-GL with the ones deduced in previous computational studies (Reinitz and Sharp 1995, Kazemian, Blatti et al. 2010) highlights several edges that previous models had failed to identify but have been corroborated by *in vivo* experiments. For example, the "TLL $\rightarrow$ *h* Stripe 6" and "TLL $\rightarrow$ *h* Stripe 7" edges in our network were suggested previously through experiments involving *tll* mutant embryos (Riddihough and Ish-Horowicz 1991, La Rosee, Hader et al. 1997), but the enhancer-based model of our previous work (Kazemian, Blatti et al. 2010) misses both of these edges. Several such examples were also noted with respect to the network reported in (Reinitz and Sharp 1995) (not shown).

An important observation from the TF-stripe networks of Fig. 4.6 is the major role played by Zelda in setting up pair rule gene expression. Recent studies have shown Zelda (zld) to be a master regulator of early embryonic development (Liang, Nien et al. 2008, Harrison, Li et al. 2011, Nien, Liang et al. 2011), and Nien et al. (Nien, Liang et al. 2011) have specifically shown the effect of Zelda knockdown on pair-rule expression. While all four genes (*eve, h, run,* and *gt*) showed severely modified expression in *zld*$^{-}$ experiments, a closer examination of Fig. 4.5 in (Nien, Liang et al. 2011) reveals specific effects that are in agreement with our TF-stripe network. For instance, the *h* gene shows complete abolishment of stripes 1, 2, 4, consistent with our predictions of direct Zelda influence on stripes 1, 2, 3, and 4 of this gene. Similarly, the most pronounced effect of Zelda knockdown on *run* expression is the abolishment of stripes 1, 2, 5, and 6, and our network predicts direct effects of Zelda of stripes 1 and 2. We are not aware of any previous computational modeling effort that predicts these specific effects of Zelda.

We should note that, our reported success in recapitulating known regulatory edges is based on our own literature survey where we have tried to be as exhaustive as possible, but admittedly we might have missed some results. As such, the high rate of recapitulated network edges is a preliminary, rather than an absolute, assessment of the accuracy of these networks.

4.2.7 Modeling cross-talk between enhancers results in aberrant expression readouts

Several studies make a case for interactions between enhancers of a gene (Kirstein, Sanz et al. 1996, Barolo and Levine 1997, Spitz, Gonzalez et al. 2003, Gonzalez, Duboule et al. 2007, Yao, Phin et al. 2008, Prazak, Fujioka et al. 2010, Montavon, Soshnikova et al. 2011), raising doubts about enhancer modularity or independence (Maeda and Karch 2011, Barolo 2012). Our experience in computational modeling of gene expression, as reported above, seems to suggest that enhancer independence is the common case. GEMSTAT-GL, which assumes independence of enhancer activities and linear aggregation of their readouts, fits expression data accurately, while GEMSTAT, which interprets all binding sites in the locus together, completely failed to fit the data. We investigated the source of this dichotomy in a systematic way, by modifying GEMSTAT-GL to allow for a limited degree of interaction (non-independence) between enhancers and noting cases where such interaction leads to a marked deterioration in model fits. We report this analysis for the enhancers of *eve, h, run*, and *gt*.

Let C_1 and C_2 be two non-overlapping enhancers (and the only two enhancers) in a locus. Let *E* denote the gene expression profile and let $G(C_i)$ denote the readout predicted by GEMSTAT for any enhancer C_i. As described in the previous sections, GEMSTAT-GL tests how well C_1 and C_2 explain *E* by computing w-PGP(*E*,

$G(C_1)+ G(C_2)$), i.e., the similarity between gene expression profile E and the integrated output of C_1 and C_2. (We ignore weights of summands here, for simplicity.) Now, let us consider any sub-segment *c* of C_2 and represent by $G(C_1, c)$ the GEMSTAT prediction on the set of binding sites in C_1 and *c* considered together. This simulates an interaction between C_1 and a part of C_2. We may now use w-PGP(E, $G(C_2) + G(C_1,c)$) as the accuracy of a model where the outputs of C_1 and C_2 are no longer independent, and in particular, the output of C_1 is shaped by contributions from a part of C_2. Let G_1 and G_2 denote two GEMSTAT-GL models (i.e., two different parameter settings) trained to optimize w-PGP(E, $G_1(C_2) + G_1(C_1)$) and w-PGP(E, $G_2(C_2) + G_2(C_1,c)$), respectively. Our goal is to find a *c* such that w-PGP(E, $G_2(C_2) + G_2(C_1,c)$) < w-PGP(E, $G_1(C_2) + G_1(C_1)$), i.e., where the model with enhancer interaction is significantly worse than the additive model. Likewise, we search for a subsegment *c* of C_1 such that w-PGP(E, $G_3(C_1) + G_3(C_2,c)$) < w-PGP(E, $G_1(C_1) + G_1(C_2)$) where G_3 is a new GEMSTAT-GL model trained to optimize w-PGP(E, $G_3(C_1) + G_3(C_2,c)$). The discovery of any such subsegment of either C_1 or C_2 will point to an *avoided interaction* between the two enhancers, i.e., a specific example in support of the enhancer independence assumed in GEMSTAT-GL.

We show in Fig. 4.7, using a heat map, the outcome of the above analysis performed on the five enhancers contributing towards the *eve* gene's expression. Rows in this heat map represent binding sites within the enhancers, and columns represent enhancers. The cell at row *i* and column *j* represents the effect (on model fits) of allowing the binding site *i* to interact with enhancer *j*. Red indicates that modeling this interaction leads to worse fits, suggesting that the interaction is avoided in reality through unknown mechanisms of enhancer independence. Green color in the heat map suggests a synergistic interaction.

Heat maps for the four genes modeled in this study (Fig. 4.7) highlighted the necessity of their enhancers to act autonomously. The many red cells indicate that such interaction must be explicitly avoided. For instance, we noted that a segment containing KR sites within the *eve* stripe 2 enhancer (Segment S_1, Fig. 4.7) adversely affects the predicted readout of the *eve* stripe 3+7 enhancer. These KR sites, when included in modeling the stripe 3+7 enhancer result in a weaker stripe 3, since the expression domain of KR covers *eve* stripe 3. A similar effect is noted for a second segment in the stripe 2 enhancer (Segment S_2, Fig. 4.7) that contains four KNI sites, which adversely influence modeling of the stripe 3+7 enhancer. Although the latter contains several KNI sites, the four additional KNI sites impart more repression than necessary and hence a deterioration in the quality of fit. (This deterioration is, however, less severe than that caused by the first segment.) These examples provide more detailed insights into why we failed in our initial attempts to model gene expression from an entire locus using GEMSTAT, where all such interactions were allowed.

4.3 Methods

Methodological details of the experiments reported in this chapter are given below.

4.3.1 Constrained parameter estimation strategy

To guard against over-fitting, we used the following model training strategy. We first trained GEMSTAT on ~40 enhancers with A/P patterned expression (Segal, Raveh-Sadka et al. 2008), while excluding enhancers of the given gene. Training on this large dataset greatly constrains the model and rules out

over-fitting. We used the parameter values thus obtained as the starting point of the parameter training procedure on regulatory sequences of the given gene. Thereafter, the training procedure was prohibited from altering any parameter's value by more than two fold from its initial value. This strategy ensured that the final model trained on the given gene is largely consistent with a model that reflects other regulatory parts of the genome.

4.3.2 Modeling a gene locus with GEMSTAT

This was performed just as any individual enhancer would be modeled by GEMSTAT. The inputs were the sequence of the locus, and the motifs and concentration profiles of the nine TFs. GEMSTAT's goal was to learn parameters such that its prediction for the readout of the entire locus matches the gene expression pattern, as quantified by the "weighted Pattern Generating Potential" (w-PGP) score described in the next paragraph. Also, since we claim (see RESULTS) that GEMSTAT is unable to model gene loci, we used an unconstrained parameter estimation strategy where the model training procedure was free to use any parameter values within a reasonable range.

4.3.3 GEMSTAT-GL model for predicting gene expression from intergenic sequence

The new quantitative model for predicting gene expression from the entire locus of a gene operates in two tiers. Recall that the inputs to the model are (i) the sequence of the locus, TF motifs, and TF concentration profiles along the A/P axis, and (ii) the gene's expression profile (assumed here to be multiple stripes along the axis). The trained model comprises (i) a set of windows (possibly of varying length, and possibly overlapping each other) in the locus, and their "window weights" (positive numbers), and (ii) values for GEMSTAT parameters reflecting TF-DNA, TF-BTM, and TF-TF interactions. The model's prediction of gene expression is the weighted sum of readouts from every window in the model, the readouts being predicted by GEMSTAT, and the weights being the window weights mentioned above.

We describe here the procedure for training the two-tiered model, given its input. The procedure learns optimal values of the GEMSTAT parameters as well as "window weight" parameters (see above) that maximize the w-PGP score between the gene expression profile and the model's prediction. A model is denoted by $M = (\boldsymbol{W}, \sigma, \theta)$ where $\boldsymbol{W}$ is a set of sequence windows from the locus, σ is the set of window weights, one for each window in $\boldsymbol{W}$, and θ is a set of GEMSTAT parameters. The model training happens in two phases. In the beginning, θ is set to GEMSTAT parameters learned from a large set of known enhancers excluding any known enhancers of the target gene.

Phase 1: In the first phase, the algorithm scans the intergenic sequence to find $N=5$ best sequence windows for each stripe in the gene expression pattern. To do so, it examines every window starting at 100 bp intervals in the locus, and of length between 500 bp and 2500 bp. (These are user-configurable parameters.) It scores every window *W* against every stripe *S* of the target gene expression, based on how well the expression read-out of *W* (predicted by GEMSTAT) fits the expression profile of *S*. The fit is quantified by the w-PGP score. At the end of this phase, the algorithm has found a set of *N* best windows for each stripe *S*, denoted by *C(S)*.

Phase 2: Next, the algorithm iteratively selects windows to include in the model, and learns their corresponding window weights. In the i^{th} iteration, it builds a model $M_i = (\boldsymbol{W}_i, \sigma_i, \theta)$ for the first i stripes of gene expression. A pseudo-code is provided next.

Given: GEMSTAT parameters θ and a candidate set of windows $C(S)$ for every stripe S.

Initialization: $\boldsymbol{W}_0$:= NULL; $BESTSCORE_0$:= 0.

For i := 1 to K (the number of stripes in gene expression pattern) do:

1. $\boldsymbol{W}_i$:= $\boldsymbol{W}_{i-1}$; $BESTSCORE_i$:= $BESTSCORE_{i-1}$;
2. For each window w in $C(S_i)$ do
    a. Define a new set W' = $\boldsymbol{W}_i$ U {w}
    b. Let σ be a set of window weights, one weight for each window in W'. Let $Score_i(\boldsymbol{W'}, \sigma)$ denote the w-PGP score that compares (i) the two-tiered model predictions using windows of $\boldsymbol{W'}$ and window weights σ, and (ii) the gene expression pattern limited to the first i stripes or expression domains. (The stripes were considered arbitrarily from anterior to posterior.)
    c. Find σ that maximizes $Score_i(\boldsymbol{W'}, \sigma)$ over all possible σ. This maximization is performed through alternating between the Simplex and the Gradient Descent algorithms for numerical optimization. Denote $\max_{\sigma} Score_i(\boldsymbol{W'}, \sigma)$ by $Score(w)$.
3. Let w^* denote the window that maximizes $Score(w)$ in the previous step.
4. If $Score(w^*)$ is greater than $BESTSCORE_i$, then
    a. $\boldsymbol{W}_i$:= $\boldsymbol{W}_i$ U {w^*}
    b. $C(S_i)$:= $C(S_i) \backslash w^*$
    c. $BESTSCORE_i$:= $Score(w^*)$
    d. Loop back to (2).

At the end of this phase, a model $M = (\boldsymbol{W}, \sigma, \theta)$ has been found for the entire expression pattern. Now, the GEMSTAT parameters θ are retrained while keeping $\boldsymbol{W}$ and σ fixed. The algorithm then loops back to Phase 1. It iterates through these two phases until a constant number N_I of iterations have been completed or the improvement in the model's w-PGP score is less than a small constant $\delta > 0$. We set $N_I = 100$ and $\delta = 10^{-4}$ for training the models in this chapter.

4.3.4 Control experiments:

(1) One of the negative control experiments involved modeling a gene's expression pattern from a randomly generated sequence of the same length as the gene locus. The random sequence was generated by independently sampling each nucleotide from a common frequency distribution. (2) Another negative control experiment involved modeling a "random" expression pattern from the sequence of a gene locus. Random expression patterns were generated based on the gene's real expression pattern, as follows. First, for any axial position, let us define the gene to be OFF if the expression value is less than 0.5 and ON otherwise. Then, for a gene G whose actual expression profile has N stripes and K axial positions where it

is ON, we defined a "random" expression profile as one where: (a) the number of stripes is a randomly chosen number between *N/2* and *N*, (b) the stripes are located randomly along the A/P axis, and (c) there are *K* data points where it is ON. (3) In the final set of negative control experiments we used a "variant" of a gene's locus, where the TF binding sites were relocated to randomly selected positions within the locus, to model the gene's expression pattern.

4.3.5 Sampling the two-tiered model

As noted above, a model is denoted by $M = (\boldsymbol{W}, \sigma, \theta)$ where $\boldsymbol{W}$ is a set of sequence windows from the locus, σ is the set of window weights, one for each window in $\boldsymbol{W}$, and θ is a set of GEMSTAT parameters. We described above a local search algorithm to find the optimal model. We also performed MCMC sampling of the space of all possible windows and window-weights, i.e., $(\boldsymbol{W}, \sigma)$ for a global examination of the expression contributions of segments in the locus.

Sample space: Each sample is an *extended* weight vector $\boldsymbol{\sigma}$ that has one real number for every possible window in the locus. Recall that this includes windows of length between 500 and 2500 (in increments of 50), with start positions that are multiples of 100 bp. Note also that any $\boldsymbol{\sigma}$ corresponds to a particular model $M_{\boldsymbol{\sigma}}$: the window set $\boldsymbol{W}$ is determined by the non-zero weights in $\boldsymbol{\sigma}$, and the GEMSTAT parameters θ are assumed fixed. The w-PGP score of model $M_{\boldsymbol{\sigma}}$ is denoted by Score($\boldsymbol{\sigma}$), and the MCMC attempts to sample $\boldsymbol{\sigma}$ with probability proportional to Score($\boldsymbol{\sigma}$).

Sampling algorithm: We used the Metropolis-Hastings algorithm to sample $\boldsymbol{\sigma}$. The allowed moves from a current sample $\boldsymbol{\sigma_i}$ are determined as follows. Let $\boldsymbol{b_i}$ be a bit vector of the same dimensionality as $\boldsymbol{\sigma_i}$ and its j^{th} bit being defined as $b_{ij} = 1$ if $\sigma_{ij} > 0$ and $b_{ij} = 0$ otherwise. That is, $\boldsymbol{b_i}$ indicates which windows have positive weights in $\boldsymbol{\sigma_i}$. The samples reachable in one move from the current sample $\boldsymbol{\sigma_i}$ (with bit vector $\boldsymbol{b_i}$) are those with bit vectors within a Hamming distance of 2 from $\boldsymbol{b_i}$. In other words, any move adds or deletes at most two windows from consideration in the first tier of the model. The proposal distribution of the Metropolis Hastings algorithm is described next. Given a current sample $\boldsymbol{\sigma_i}$ (with bit vector $\boldsymbol{b_i}$), we choose two bits at random and toggle each bit with probability 1/2. This samples a bit vector $\boldsymbol{b_j}$ that is (a) identical to $\boldsymbol{b_i}$ with probability ¼, (b) 1 Hamming distance from $\boldsymbol{b_i}$ with probability ½, and (c) 2 Hamming distance from $\boldsymbol{b_i}$ with probability ¼. All bit vectors with a particular Hamming distance are equally likely. There are $L = |\, b_i \,|$ of these at Hamming distance 1, and $\binom{L}{2}$ of these at Hamming distance 2. The newly sampled bit vector $\boldsymbol{b_j}$ is then used as the "shape vector" of a Dirichlet distribution, from which a probability vector is sampled. This is the newly sampled weight vector $\boldsymbol{\sigma_j}$. As prescribed by the Metropolis Hastings algorithm, this proposed sample $\boldsymbol{\sigma_j}$ is then accepted with probability min(1, Score(σ_j)/Score(σ_i)).

4.3.6 Constructing heatmaps to study enhancer interactions

Our goal was to probe potential interactions between binding sites from two different enhancers. In particular, we wanted to determine if interpreting the sites of one enhancer together with sites from another enhancer leads to better model predictions than the baseline of GEMSTAT-GL where each enhancer is interpreted independently. A natural way to represent such potential interactions is with a

hypergraph. Hypergraphs generalize the concept of graphs by allowing each edge (called "hyperedge") to represent a relationship shared among more than two nodes. In our formulation, every binding site in every enhancer is a node in a hypergraph, and any subset of sites from two different enhancers defines a hyperedge. The evidence in favor of that subset of sites being interpreted together, as if they were sites in the same enhancer, is the weight of the hyperedge. Such weights can be negative also, indicating that the particular subset of sites if interpreted together will make model predictions worse. We limited our attention to hyperedges defined by including (a) all sites of one enhancer and (b) sites within a sub-segment of a different enhancer, thus simplifying the space of enhancer interactions considered. Our model-based predictions of potential interactions (or avoidance of interactions) can be captured by this weighted hypergraph. However, a hypergraph is hard to visualize and less likely to lead to biological insights via direct examination. We therefore mapped the constructed hypergraph to a weighted graph where the weight of every edge represents the effect of allowing interaction between the two binding sites that the edge represents. Visualization of the edge weights of this graph through heatmaps then revealed how any binding site could affect the readout of any enhancer in our model.

Hypergraph construction: For a gene *g*, suppose the GEMSTAT-GL model selects *n* contributing enhancers $C_1, C_2, \ldots, C_n$. Let $SITES(C_i)$ denote the set of TF binding sites in enhancer C_i. Then, for every binding site in every set $SITES(C_i)$, we include one node in a hypergraph. There are two types of hyperedges in our hypergraph. First, every subset of $SITES(C_i)$ constitutes one hyperedge, and every such hyperedge was assigned a weight of zero. Each of the remaining hyperedges represents a collection of binding sites from two different enhancers, and was constructed as follows. Let e_h denote a hyperedge that consists of binding sites from enhancers C_1 and C_2. Then the hyperedge e_h would include all the binding sites of one enhancer (say, C_1) and between one and five contiguous binding sites of the other enhancer (C_2 in this case). For each hyperedge e_h constructed in this way, we optimized a new GEMSTAT-GL model where the contributing enhancers are $C_2, \ldots C_n$, as well as the newly constructed set of sites e_h treated as an "enhancer". The difference between the w-PGP score of this new model and the original model learned for gene *g* was then assigned as the weight of e_h.

Mapping hypergraph to graph: A graph was constructed with the same nodes as that in the hypergraph, with an edge for each pair of nodes. The weight of an edge was computed by averaging the weight of every hyperedge where the corresponding pair of nodes appeared. This approach of approximating a hypergraph through a graph was discussed in detail in (Agarwal, Jongwoo et al. 2005). By construction, this graph has the property that the edge between node *i* and node *j* has the same weight for all nodes *j* corresponding to sites in the same enhancer.

4.4 Discussion

We have presented for the first time a quantitative model that relates gene expression to the sequence of an entire gene locus, using information on the trans-regulatory context (TF concentrations). We started by showing that the thermodynamics-based model "GEMSTAT" accurately models individual enhancer readouts, but fails to model the entire locus. We then performed a series of tests where we changed the way the GEMSTAT model was applied to the locus, all of which resulted in failure. We developed a new model called GEMSTAT-GL where the expression readout of the locus is two-tiered: sites within each

enhancer act together to produce that enhancer's contribution, and contributions from multiple enhancers are aggregated to produce the gene's expression pattern. This model shows very good fits to the data (Fig. 4.3 and Fig. 4.4) for the 27 genes studied here, and most remarkably for the complex, seven-stripe patterns of *eve*, *h*, and *run*. The process of training the model on a gene locus automatically predicts enhancers in that locus, without relying on chromatin accessibility data, and makes accurate assignments of regulatory activity to each of the predicted enhancers. We will make available, upon publication, a general-purpose implementation of the GEMSTAT-GL model that may be applied to any gene for which the relevant inputs (TFs, TF motifs, TF concentrations) and output (gene expression) are known. The implementation also allows users to include chromatin accessibility data as a filter on the locus being modeled.

We note that the GEMSTAT-GL model, as presented here, is given an intergenic sequence and its expression readout, and it finds a plausible explanation of whether and how the sequence could drive that expression. As such, it can be applied, in principle, to any of the thousands of genes whose embryonic expression patterns are known from *in situ* hybridization assays (Tomancak, Berman et al. 2007). Once trained, the model reveals the cis-regulatory architecture of the locus (locations and readouts of individual enhancers), and can predict the effects of perturbations in cis (sequence) or trans (TF concentration).

However, the model cannot currently be used to predict the expression readout of a gene locus from sequence only. This is because the locations of contributing segments in the locus are free parameters of the model and can be learnt only if the gene expression readout is known. Thus, the model performance reported here refers only to "training data accuracy", and leaves open the possibility of over-fitting. However, the model training failed on a variety of different "negative control" tests, where there was no link between the given sequence and expression, thus addressing concerns of over-fitting. We expect future work to address the current limitation that prevents the new model from a full-fledged application to the genome. One way this may be achieved is through intelligent use of accessibility and chromatin state information (Ernst and Kellis 2010, Kharchenko, Alekseyenko et al. 2011) from the locus when selecting segments that contribute to gene expression.

Another potential limitation of this work is its reliance on prior knowledge of the TFs relevant to the regulatory system being studied (the A/P patterning system here). Ideally, the model should be able to automatically identify the TFs that are needed to explain the data, but this ability was not tested in this work. In a separate work (Samee and Sinha 2013), we address the question of systematically identifying the TFs to use when modeling enhancers using GEMSTAT.

A basic principle underlying GEMSTAT-GL is the modular view of the gene locus' readout, which holds that individual enhancers drive discrete aspects (e.g., one or two stripes) of the gene's expression pattern, through combinatorial action of the binding sites within them, and the overall gene expression pattern results from a superposition of these separate enhancer readouts. Our tests showed that a model that violates this modular view and instead interprets all binding sites in the locus as acting together is unlikely to fit the data. In other words, the rules for interpreting the set of sites across all enhancers are not the same as the rules that apply to sites within an enhancer. The final subsection of RESULTS provides details of this principle in action: the different enhancers have the potential to interfere with each other, i.e., if

some sites in one enhancer, say S_i, are interpreted together with sites of another enhancer, say S_j, the combined readout may be different from the readout of S_j itself.

Another defining aspect of our model is the use of a "weighted sum" as the aggregator of multiple enhancer readouts. We note that the weights assigned by the model to different contributing segments (enhancers) are comparable to each other, and that a simple unweighted sum captures the seven stripe pattern of gene expression qualitatively, but fails to capture the "valley" between stripes 2 and 3 for *eve* and between stripes 4 and 5 for *run*; whereas the prediction for *h* remains relatively unaffected. Thus, the use of non-uniform weights may be a way for our model to correct for inaccuracies of the GEMSTAT model in predicting enhancer readout, especially at stripe borders. These weights need not be a reflection of any fundamental biochemical preference for one enhancer over another.

One may speculate on biochemical mechanisms that implement the two-tiered readout of the regulatory information at the locus, and the additive aggregator function. An obvious possibility is that each contributing segment interacts with the promoter separately, as shown in Fig. 4.8-A. In the example shown, there are two enhancers and three possible configurations of enhancer-promoter interaction. The "Boltzmann weight" of each configuration is assumed to depend only on the enhancer interacting with the promoter in that configuration. Let these weights be 1, η_B and η_A for the configurations at the top, middle and bottom respectively. Assuming that gene expression 'E' is proportional to the total probability 'p_{locus}' of configurations with any enhancer-promoter interaction, we get:

$$E \propto p_{locus} = \frac{\eta_A + \eta_B}{\eta_A + \eta_B + 1}$$

First, let us consider a trans-regulatory context where one of the enhancer (say A) drives expression and the other (say B) does not. This can be formulated as: $\frac{\eta_A}{\eta_A+1} = p$ and $\frac{\eta_B}{\eta_B+1} \ll p$

Under these conditions, we get $p_{locus} \approx p$, i.e., the contributions of the two enhancers add up to produce the expression driven by the locus. Thus, if a gene is under the control of multiple enhancers and if a single enhancer dominates all others in any particular trans-regulatory context (position along the A/P axis, for pair rule genes), we expect the combined readout of the multiple enhancers to be a sum of their individual readouts.

Now consider a trans-regulatory context where both enhancers A and B (of Fig. 4.8-A) have comparable outputs. We may formulate this as:

$$\frac{\eta_A}{\eta_A + 1} = \frac{\eta_B}{\eta_B + 1} = p$$

It is easily shown that in this case

$$p_{locus} = \frac{2p}{1 + p}$$

We plot this function, representing the combined readout of the locus, in Fig. 4.8-B. For small values of *p* (< 0.2), this function is reasonably approximated by 2*p*, indicating that the enhancer contributions add up. Note that a value of *p* = 0.2 does not necessarily mean low gene expression; under the Shea & Ackers theory, expression levels are only *proportional* to *p* as defined here. For larger values of *p*, we see that p_{locus} is better approximated by the function 1 – (1-p)^2, which represents the model of enhancer synergy proposed by Perry et al. (Perry, Boettiger et al. 2011). In this case, the separate readouts of enhancers do not combine additively. Another scenario in which additivity is not expected is where multiple enhancers can interact simultaneously with the promoter, as is the case in the "long range dominant repression" model of Perry et al. (Perry, Boettiger et al. 2011). We explicitly prohibited such a configuration in the model of Fig. 4.8-A.

In light of the simplistic arguments presented above, we suggest that the model illustrated in Fig. 4.8-A, with the strength of each enhancer-promoter interaction being unaffected by other enhancers in the locus, as a mechanistic basis of the GEMSTAT-GL model. Additivity of enhancer contributions in any given trans-regulatory context can be explained by this model as arising out of (a) one enhancer's contribution dominating all others or (b) each enhancer's contribution being at a relatively low level, i.e., the probability *p* defined above being not close to 1.

4.4.1 A note on parameter estimation for the locus-level modeling problem

A locus-level model of gene expression requires more precision than an enhancer-level model. The success of an enhancer-level model is typically assessed from its precision in modeling the position of the peaks of the expression domains driven by an enhancer. Consequently, the most successful enhancer-level models produce qualitatively accurate expression patterns for each enhancer but may not capture the peak amplitude of expression domains correctly. That is, relative peak amplitudes of readouts from two enhancers are often inconsistent with model predictions. Another type of imprecision noted in enhancer-level models is the inability to predict the sharp boundaries of expression domains. A locus-level model cannot afford to tolerate such imprecision, especially when it is applied to model complex multi-stripe expression patterns. The two weaknesses of enhancer-level model fits mentioned above can cause our locus-level model to predict qualitatively inaccurate expression patterns (e.g., miss an inter-stripe boundary), and are likely to lead to false regulatory sequence discovery and wrong inference about the roles of TFs. At the same time, the quantitative imprecision in predicted enhancer readouts may be unavoidable at this time due to fundamental limitations of the thermodynamic model, e.g., biochemical mechanisms that are not modeled.

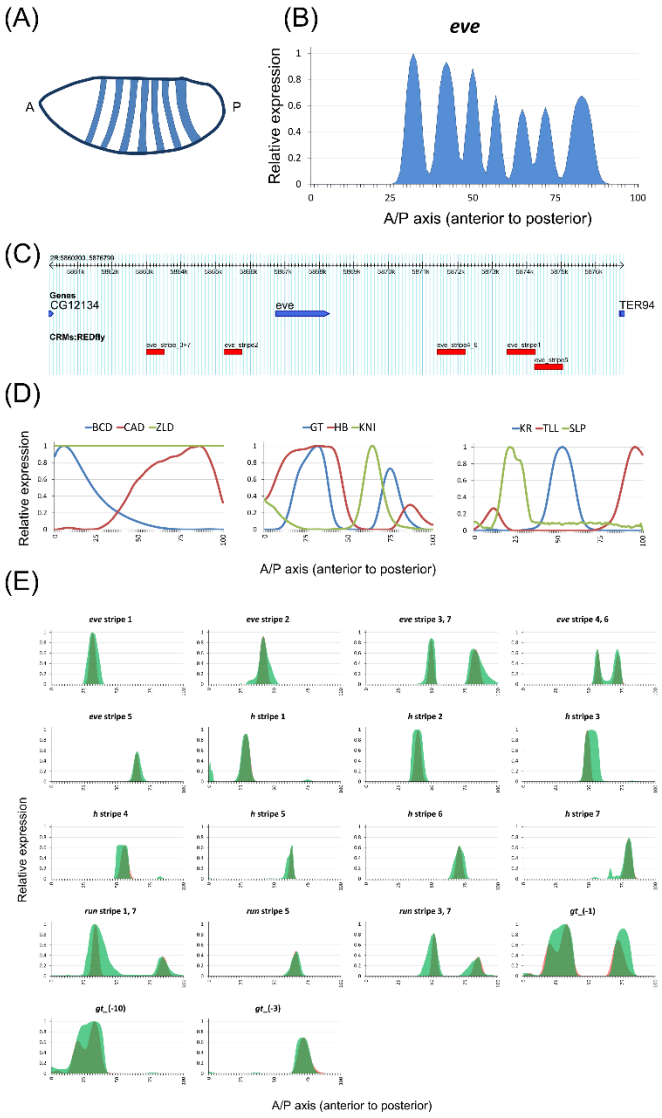
Our strategy of optimizing the thermodynamic parameters for each gene separately was a pragmatic decision made to compensate for the minor inaccuracies of enhancer-level modeling. When GEMSTAT-GL was optimized without re-training the thermodynamic parameters (thus, the locations and the weights of the windows were the only free parameters in the model), it could still capture the correct locations for five of the seven stripes of *eve* expression but suffered severely in terms of modeling the inter-stripe valleys. Thus, fitting the thermodynamic parameters in a locus-specific manner helps GEMSTAT-GL to achieve the desired accuracy. It is plausible that this strategy might lead to over-fit GEMSTAT-GL for the single intergenic locus being modeled. This is why we performed four different types of negative controls,

to demonstrate that the constraints imposed on the parameters during model optimization are strongly guarding us against over-fitting the model for any specific locus.

In a recent study (Kim, Martinez et al. 2013), Kim et al. trained thermodynamics-based models on a collection of *eve* enhancers in order to provide deeper insights into combinatorial cis-regulatory logic, which, as they pointed out, is a pre-requisite for locus-level modeling of gene expression. Among other findings, they reported a model that predicts *eve* stripes 2, 3, and 7 from the sequence upstream of the gene, and a different model (i.e., different parameter settings) that predicts stripes 4, 5, and 6 from the sequence downstream of the gene. Their results, in addition to providing insights about functioning of enhancers, highlight the difficulty of modeling the readout of an entire gene locus using pre-determined parameters, even when the models are accurate at the enhancer level. This agrees with our own view mentioned above, and suggests that fitting thermodynamic parameters for individual loci, with appropriate constraints, is a necessary step at the current stage of computational modeling of gene expression from the locus.

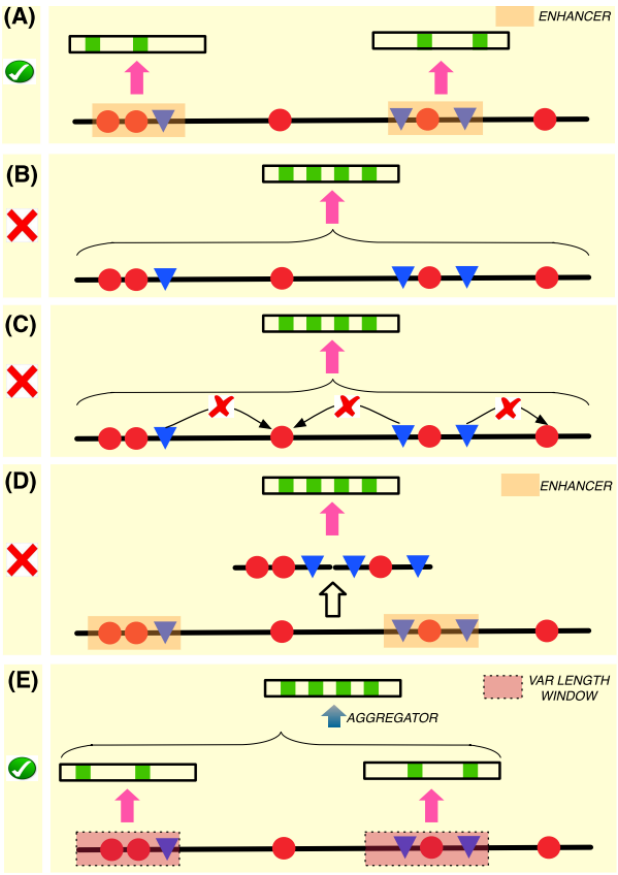
4.5 Figures

Figure 4.1: Examples of complex expression patterns and GEMSTAT's performance on the individual enhancers of the corresponding genes.



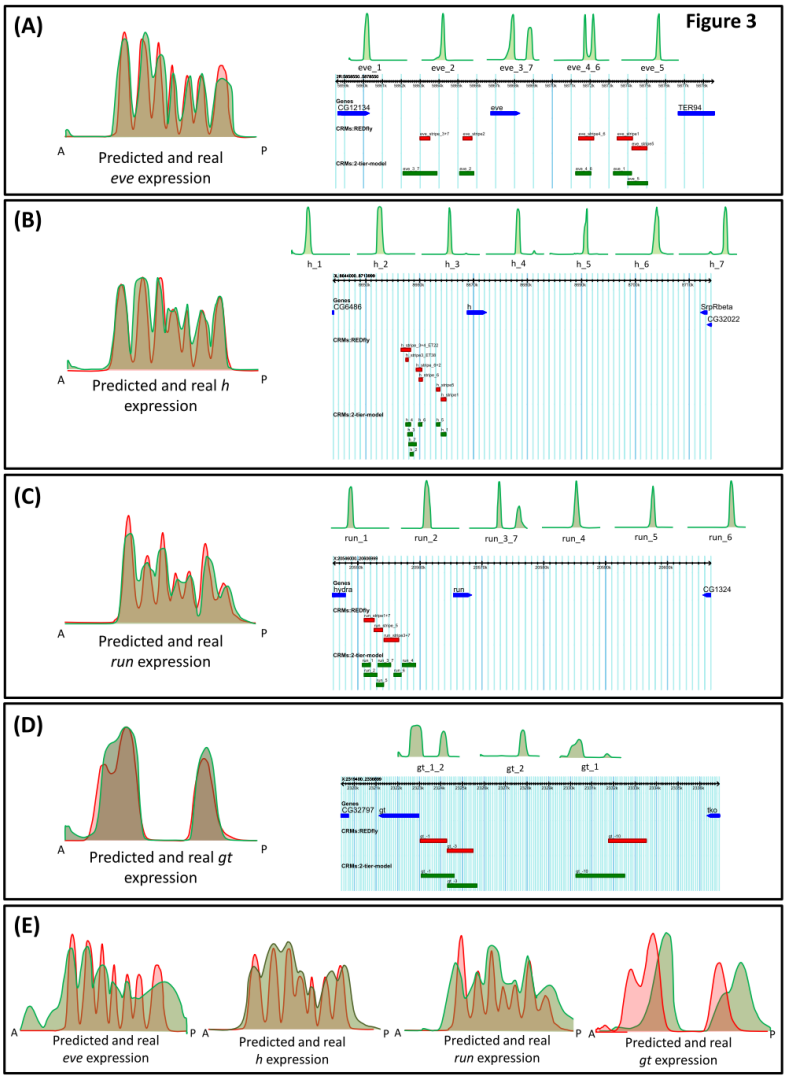
(A) Schematic of expression pattern of the pair-rule gene *even-skipped* (*eve*) in *D. melanogaster* embryo. 'A' and 'P' denote the anterior and the posterior ends of the embryo, respectively. (B) Quantitative profile of *eve* gene expression along the anterior-posterior axis of the embryo. (C) Genome Browser view of the five distinct enhancer elements that drive *eve* gene expression; each enhancer's name denotes the specific stripe(s) of gene expression that it drives. The entire locus is 17 Kbp long. (D) Concentration profiles along the anterior-posterior axis, for the nine TFs used to model the expression patterns of the genes *eve*, *h*, *run*, and *gt*. (E) Real (red) and GEMSTAT-predicted (green) expression profiles along the A/P axis for the known enhancers of *eve*, *h*, *run,* and *gt*.

Figure 4.2: Systematic application of increasingly complex models to compute gene locus' readout.



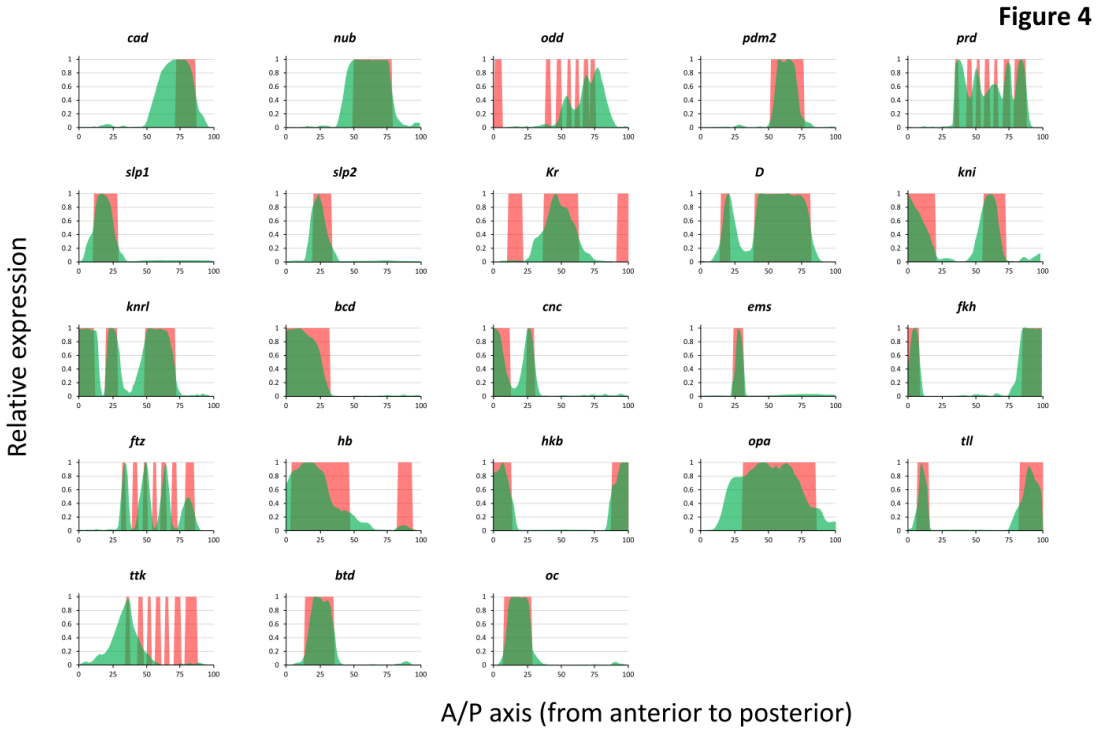
A hypothetical example illustrating the different attempts at developing a locus-level model of gene expression. The hypothetical gene here is expressed in four stripes - shown in panels (B)-(E) as green stripes within a rectangle. The thick black line near the base of each panel denotes the locus; red circles and blue triangles denote activator and repressor binding sites within the locus, respectively. The bold pink arrow indicates GEMSTAT prediction of an expression readout on a given segment. (A) GEMSTAT accurately models the 2-striped expression patterns driven by "known" enhancers for this hypothetical gene. (B) GEMSTAT fails to model the 4-striped readout of entire locus in the "Direct Interaction" mode. (C) GEMSTAT fails to model the locus readout in the "Short Range Repression" mode (quenching effect of repressor sites is shown using the red 'X' marks on arrows connecting repressor sites to nearby activator sites). (D) GEMSTAT also fails to model the gene's 4-striped expression from the concatenation of its two "known" enhancers. (E) A two-tiered model, that first selects a handful of variable-length windows (putative enhancers) from the locus and then takes a weighted summation of the GEMSTAT-predicted readouts of those windows to model gene expression. This model produces accurate fits.

Figure 4.3: Predictions of the GEMSTAT-GL



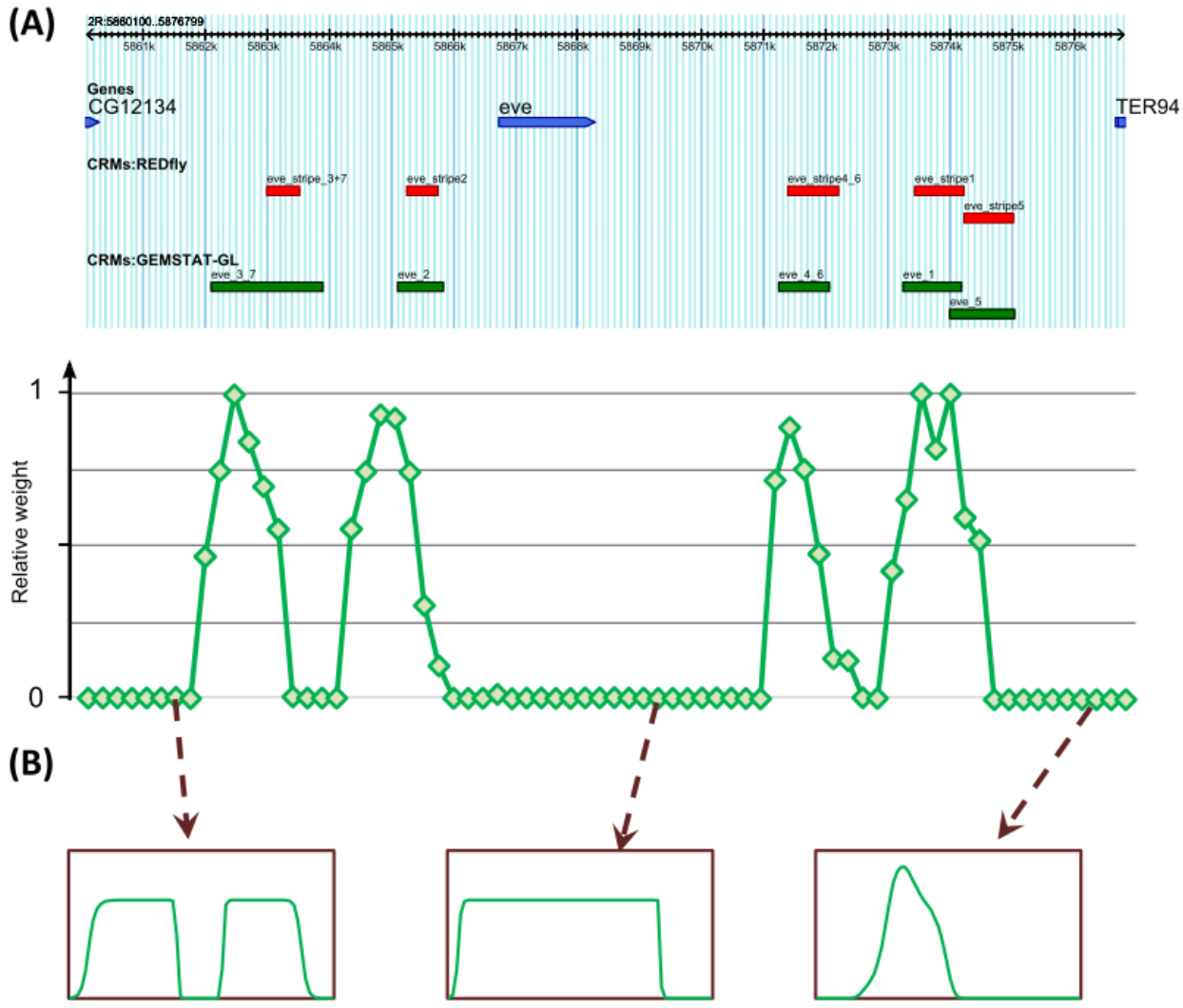
(A) Results of applying GEMSTAT-GL on the intergenic region of *eve* in *D. melanogaster*. Left panel shows the real (red) and predicted (green) expression profiles along the A/P axis. Right panel shows the locations of selected windows (green boxes) in the locus and their predicted expression patterns (top), along with locations of known *eve* enhancers (red boxes). (B), (C), and (D), same information for *h, run,* and *gt*, respectively. (E) Expression patterns modeled by GEMSTAT-GL from the intergenic regions of *eve*, *h*, *run,* and *gt* in the *D. pseudoobscura* genome.

Figure 4.4: GEMSTAT-GL on additional 23 genes



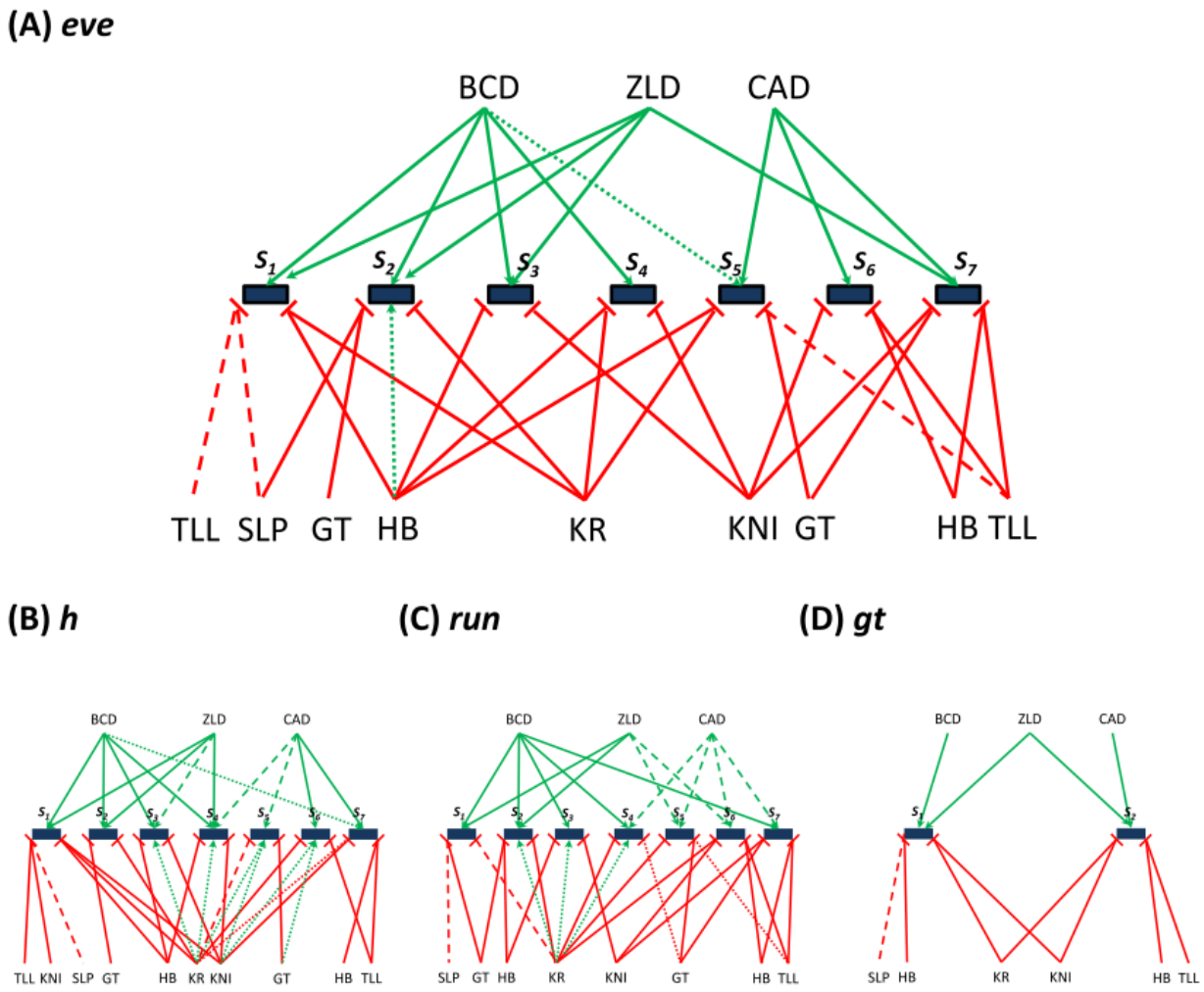
Results of fitting the GEMSTAT-GL model on the intergenic locus of 23 additional genes studied in (Kazemian, Blatti et al. 2010). Quantitative data on target expression patterns were obtained from the companion website of the same study, and were originally derived from *in situ* expression images at the FlyExpress (Kumar, Konikoff et al. 2011) database. For each gene, the red and the green plots represent the target (real) and the modeled expression patterns, respectively.

Figure 4.5: Discovering the regulatory architecture of a gene locus



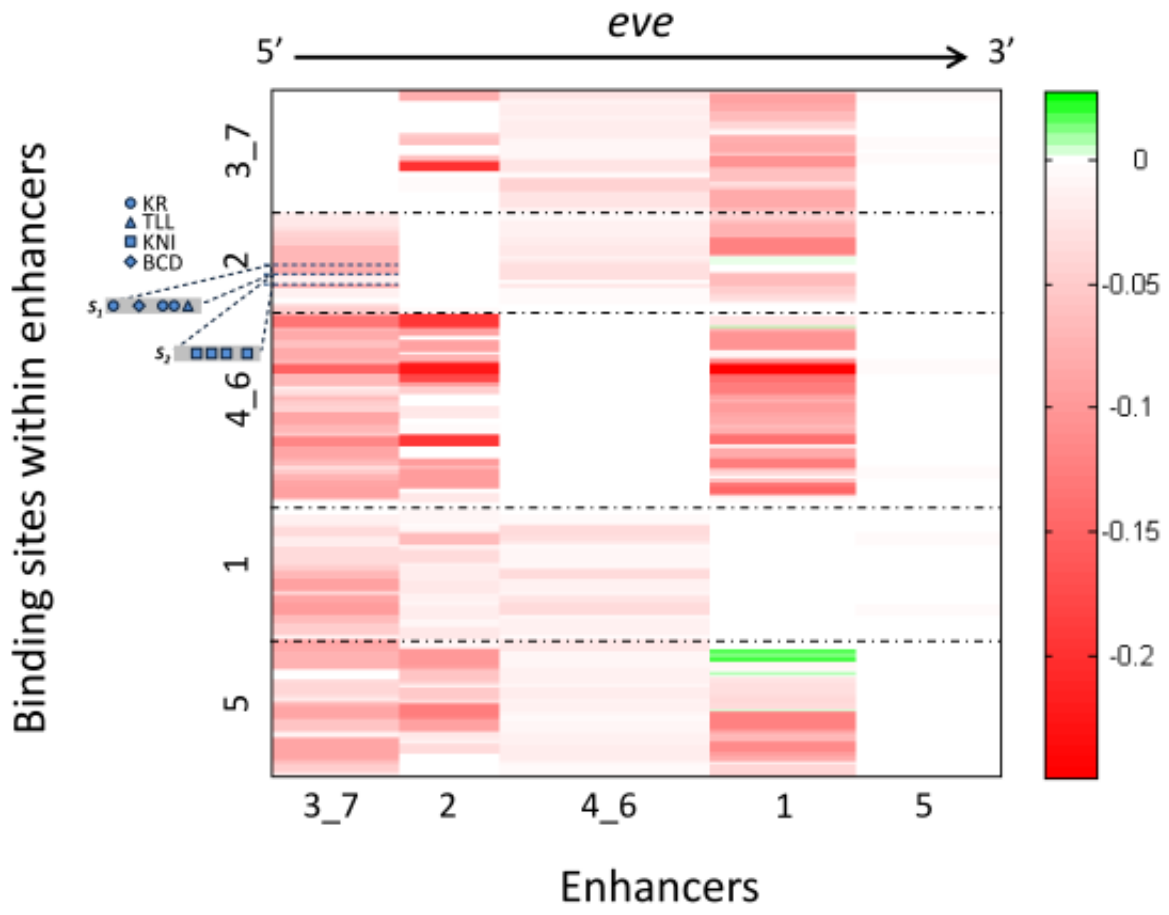
Outcome of MCMC sampling to reveal the cis-regulatory architecture of *eve* intergenic region. (A) Top panel shows the *eve* intergenic locus along with the known enhancers of *eve* and windows selected by GEMSTAT-GL to model *eve* expression pattern. Bottom panel shows the average weight of segments in the locus as estimated by MCMC sampling. The horizontal axis of the bottom panel spans the *eve* locus; green diamonds in the plot represent the starting positions of the sequence segments that comprise the MCMC samples (segments corresponding to two different green diamonds might therefore differ in length). The vertical axis denotes the average weight (on a relative scale between 0 and 1) that each segment received over 50,000 samples. (B) Predicted readouts of three zero-weight segments that could have an irreconcilable effect on the gene expression pattern, and were not selected by the two-tiered model.

Figure 4.6: Discovering TF-stripe networks



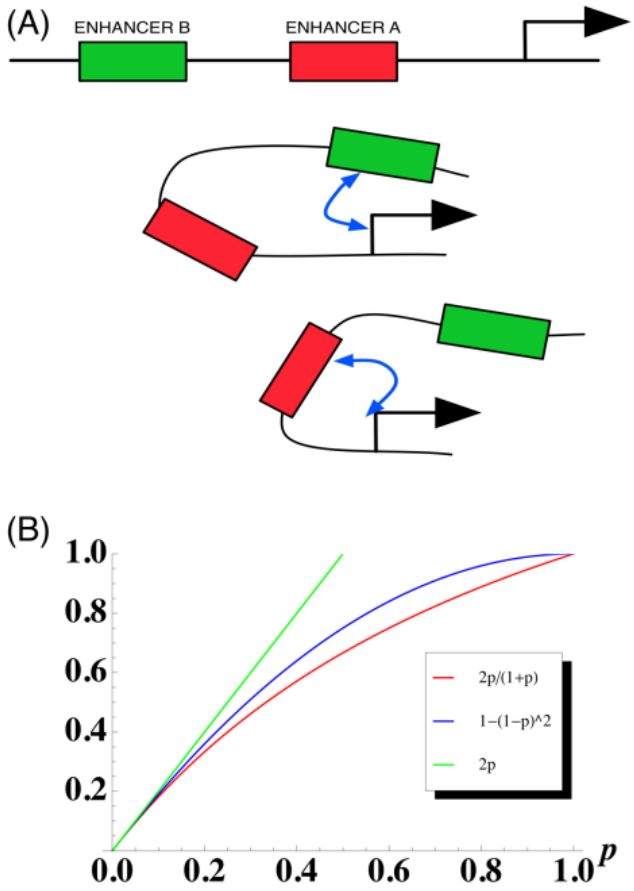
(A)-(D) Networks showing regulatory influences of TFs on individual stripes of *eve*, *h*, *run*, and *gt*, respectively. Red edges denote repressive and green edges denote activating role of the corresponding TF. Solid edges denote predicted influences that are already known in the literature. Edges with large dashes denote predicted influences that were not reported in the literature before (false positive or novel predictions), while edges with small dashes denote predicted influences already known in literature but missed by our model (false negatives).

Figure 4.7: Testing for interactions between enhancers.



A heatmap visualization of the changes in GEMSTAT-GL's goodness-of-fit owing to interactions between the enhancers selected for the *eve* gene. The heatmap has 5 columns and N_{eve} rows, where N_{eve} denotes the total number of binding sites in the five *eve* enhancers. Each row in the heatmap represents a binding site; the ordering of the rows, from top to bottom, reflects the 5' to 3' order of the respective binding sites in the locus. Each horizontal dot-dash line demarcates binding sites from two different enhancers. Each column in the heatmap represents an enhancer; the columns are ordered, from left to right, according to the 5' to 3' order of the corresponding enhancers in the locus. The cell at row *i* and column *j* represents, on a green-to-red color scale (green: high, red: low), the effect of allowing the binding site *i* to interact with enhancer *j*. This effect quantifies how the goodness-of-fit improves (green) or decreases (red) when interactions are allowed (see Methods for details). Two segments S_1 and S_2 within the *eve* stripe 2 enhancer are shown on the left of the heatmap, along with their constituent binding sites for TFs KR, TLL, KNI, BCD. Each of these segments has binding sites that, when allowed to interact with the eve_3_7 enhancer, result in poorer fits.

Figure 4.8: Molecular principles underlying the GEMSTAT-GL model



(A) A gene locus with two enhancers (A and B) can be in one of three different configurations of enhancer-promoter interaction: (top) neither enhancer interacts with promoter, (middle) only B interacts and (bottom) only A interacts. In a configuration where A interacts with promoter, B does not interact, and vice versa. (B) Combining contributions from two enhancers. If each enhancer's contribution is given by the gene expression probability *p* due to that enhancer, the combined contribution of the two enhancers (assuming independent interactions with the promoter) is 2p/(1+p), plotted in red. For small values of p, this is well approximated by 2p (green), the sum of their contributions. For larger values of *p*, a better approximation is provided by the function $1-(1-p)^2$, in blue.

Chapter 5
Thermodynamic Modeling of Fused Enhancers Reveals Novel Mechanism of Enhancer Readout

5.1 Introduction

An important goal in regulatory genomics is to understand the logic that specifies expression readout of *enhancers* – sequences that determine the expression level of an associated gene as the cellular environment varies, thus shaping the spatio-temporal expression 'pattern' of that target gene (Levo and Segal 2014). At the molecular level, the expression pattern driven by an enhancer (its readout) is realized, in major part, through the regulatory effects of transcription factor (TF) molecules bound to their cognate sites in the enhancer (Yanez-Cuna, Kvon et al. 2013). Enhancers are assumed to function independent of their genomic context (Arnold, Gerlach et al. 2013), hence their alternative designation as cis-regulatory 'modules'. Mechanistically, this implies that when TFs bound to an enhancer exert their regulatory effects on the target gene, TFs bound elsewhere do not exert any regulatory effect on that gene (Fig. 5.1-A). Note that TFs bound outside a particular enhancer can and do exert regulatory effects on the target gene, but such effects are assumed to be irrelevant to understanding and modeling the regulatory events within the enhancer and its modular readout.

Going beyond the above assumption of enhancer modularity, conventional models of enhancer readout ('enhancer-level' models) also assume that TFs bound to sites within an enhancer exert their regulatory effects simultaneously (Ay and Arnosti 2011). This additional assumption and the success of corresponding models suggests that to understand an enhancer's readout one need not be concerned about smaller segments within the enhancer functioning independently from the rest of the enhancer, just as the enhancer functions independently from the sequence context outside it. This assumption is not obvious, as the same cannot be said of the entire regulatory region ("locus") of a gene: as noted above, one does not assume that TFs bound to sites across the locus exert their regulatory influence simultaneously, rather one must delineate the locus as a collection of distinct enhancers with independently computed readouts (Buchler, Gerland et al. 2003). For instance, we showed previously (Samee and Sinha 2014) that to model a gene's expression from its regulatory region ('locus') one must consider a locus as consisting of independently functioning sub-segments (Fig. 5.1-B) whose regulatory effects are insulated from each other and are aggregated at the gene-level. This view of cis-regulatory architecture is in line with the emerging role of higher order chromatin structures in determining gene expression (Shlyueva, Stampfel et al. 2014).

In short, there seem to be two different sets of rules for interpreting a regulatory sequence, one set applicable within enhancers and a different set at the level of the entire locus. The distinction appears to be generally based on the length of the sequence, with enhancers typically being ~1-2 Kbp long (Berman, Pfeiffer et al. 2004) while the locus is often tens or even hundreds of Kbp in length. However, this partitioning of cis-regulatory length scales into two regimes (subject to interpretation by an enhancer-level model versus a locus-level model) is somewhat arbitrary, and raises several related and significant

questions. What length, if any, acts as the boundary separating the two regimes? What is the mechanistic origin of the dichotomy? Is it possible, for instance, that an enhancer-length sequence (~1 Kbp long) contains multiple independent segments that must be interpreted separately? We investigate this last question here. The question challenges our current understanding of sequence-mediated regulatory logic. The two models may lead to qualitatively different readouts from the same sequence, with one model predicting the gene to be highly expressed and the other predicting no expression, as we illustrate informally in Fig. 5.1-C and demonstrate formally in the following sections. If an enhancer may indeed be further delineated into smaller modular segments, it would reveal an additional layer of complexity in the *cis*-regulatory code, that must be systematically investigated and ultimately incorporated into quantitative models of gene expression and predictions of non-coding mutations (Yanez-Cuna, Kvon et al. 2013, Weingarten-Gabbay and Segal 2014).

It is non-trivial to answer this question since one needs a carefully chosen data set that can challenge the current enhancer-level models: these models have already proven sufficient in explaining readouts of experimentally characterized enhancers (Segal, Raveh-Sadka et al. 2008, He, Samee et al. 2010, Kazemian, Blatti et al. 2010, Samee and Sinha 2013) and several libraries of short (~100 bps) sequences generated by random insertion of TF sites (Gertz, Siggia et al. 2009, Kwasnieski, Mogno et al. 2012). However, a stated tendency towards defining minimal functional enhancers and the resulting ascertainment bias in current collections of *bona fide* enhancers means that these may not offer the opportunity to test the two contradictory models mentioned above. A possible solution is to test artificial enhancer-length sequences of 1-2 Kbp length that have synthesized by concatenating two shorter enhancers, and asking if their readout follows the enhancer-level model or the locus-level model.

In this work, we model a novel data set (Lydiard-Martin *et al*., unpublished) that was constructed specifically to challenge the conventional models of enhancer readout. The data set includes six artificial constructs created by fusing two well-studied enhancers of the Drosophila even-skipped (eve) gene in different ways. The two constituent enhancers, namely 'eve_3/7' and 'eve_4/6', drive four of the seven stripes where eve is expressed in pre-cellular stage Drosophila embryo (Fig. 5.1-D). The fused constructs are all ~1.4 kbps in length, comparable to the average length of other developmental enhancers in Drosophila (Fig. 5.1-E), and they manifest patterned readout when placed next to a reporter gene (Fig. 5.1-F). The two original enhancers (eve_3/7 and eve_4/6) contain binding sites for the same nine TFs, but their main difference is in the affinity of the sites they harbor for two of the repressor TFs, namely Hunchback (Hb) and Knirps (Kni) (Fig. 5.1-G). Due to this special property (not shared with a previous study of fused enhancers (Small, Arnosti et al. 1993)), it is not clear what the readout of a sequence that fuses the eve_3/7 and the eve_4/6 enhancers should be. If an enhancer-level model (where all bound TFs act simultaneously) is operational, one would expect two broad stripes as the readout. If a locus-level model is appropriate for the fused constructs, with the two constituent enhancers as the two independent segments, then one would expect four stripes. A third possibility is that of a locus-level model with a different delineation of independent segments, in which case it is not possible to intuitively predict the readouts. This data set therefore provides an opportunity to test the two alternative models of *cis*-regulatory organization at the length scale of typical enhancers. Details of creating the data set are provided in (Lydiard-Martin *et al*., in preparation); here we attempt to understand the mechanism leading

to the observed readouts of these constructs, and to answer the central question raised above. Computational modeling is necessary to rigorously discern which of the scenarios/models outlined above is most supported by the data.

We approached this data set by systematically fitting different state-of-the-art thermodynamics-based sequence-to-expression models to it. We found that GEMSTAT – a mechanistically rich enhancer-level model (He, Samee et al. 2010) – fits the readouts of the individual enhancers of *eve* with high accuracy, but fails completely to model the fused constructs even though they are similar in length to typical enhancers. Given the richness of the GEMSTAT model and the flexibility we allowed the model (see Methods), we interpreted this as insufficiency of the current enhancer-level models and attempted next to fit the data set using GEMSTAT-GL – a locus-level model (Samee and Sinha 2014). GEMSTAT-GL models the readout of a sequence as a weighted sum of the readouts of multiple independently functioning segments (within the sequence), which the model discovers automatically in the course of data-fitting. In light of the additional flexibility (additional parameters) allowed in GEMSTAT-GL we adopted a much more constrained strategy for model fitting (see Methods). We found that for each of the fused constructs GEMSTAT-GL selected independent regulatory segments whose readouts could be linearly aggregated to fit the construct's readout with much higher accuracy, thus providing a resolution to the motivating question of this study.

Our observation of multiple independently interpreted segments within an enhancer-like sequence suggests some form of localization of regulatory effects of DNA-bound TFs. One well-studied mechanism that results in such localization is that of "short-range repression" (SRR), whereby a bound TF molecule exerts its repressive effect within a short range (~100 bp). We investigated the SRR mechanism within our modeling framework and noted that it improves fits compared to the baseline enhancer-level model but fails to capture several salient features of data that GEMSTAT-GL was able to model. In summary, results of modeling a unique data set strongly argues that even enhancer-length sequences might work through independently functioning smaller segments present within the sequence.

5.2 Results

5.2.1 An enhancer-level model explains the readouts of the *even-skipped* enhancers but fails to explain readouts of fused constructs

We modeled the sequence-to-expression relationship of six artificial constructs that have been tested experimentally through reporter constructs in the early *Drosophila* embryo. As explained above, each artificial construct is a concatenation or 'fusion' of two well-studied enhancers 'eve_3/7' and 'eve_4/6', and has length comparable to a typical developmental enhancer. To test whether the assumptions underlying enhancer-level models hold for this data set, we fit a state of the art enhancer-level model called GEMSTAT (He, Samee et al. 2010) individually on each artificial construct to fit the corresponding readout (see Methods). The GEMSTAT model has been shown previously to produce fairly accurate fits to the readouts of ~40 enhancers involved in anterior-posterior (A/P) patterning of the Drosophila embryo and particularly for enhancers of the *even-skipped* (*eve*) gene (He, Samee et al. 2010, Samee and Sinha 2014); the two fused enhancers are known to individually regulate aspects of the endogenous expression

of this gene. We repeated the same exercise here, as shown in Fig. 5.2, to fit each fused construct's readout as a function of its sequence. Since our aim here was to test whether any parameterization of GEMSTAT can fit the readout of any artificial construct, we adopted this approach of fitting the model individually on each construct. This approach might have led to overfitting the data, but also makes it unlikely that failure to fit the data will be due to technical limitations such as optimization over a large parameter space.

As part of the modeling exercise, we first randomly sampled ~1 million models from the parameter space. Then we considered each of the top 1000 models from the sampled collection, one at a time, as the initial parameterization of the GEMSTAT model and re-estimated parameters to optimize the model for the specific construct under consideration (see Methods). Although this setup provided considerable flexibility to fit the constructs' readouts (in comparison to a typical enhancer modeling setup), GEMSTAT failed to model any construct's readout with satisfactory accuracy, as shown in Fig. 5.2. In particular, the model even after optimization could only produce two broad stripes for each construct, and failed completely to capture the salient features (see below) of each construct's readout.

5.2.2 A locus-level model can explain readouts of the artificial constructs by identifying independent regulatory segments within each construct

The failure of the enhancer-level model (GEMSTAT), as noted above, is very similar to what we had experienced in our past work on modeling the expression pattern of a gene from its entire locus (the intergenic region bounded by its neighboring genes on either side). There, we found that for several multi-stripe genes (namely, *even-skipped*, *hairy*, and *runt*), the GEMSTAT model predicted a broad domain of expression spanning all or most of the multiple stripes, failing to capture their striped patterns of expression. This led us to develop a new model called GEMSTAT-GL (GEMSTAT for Gene Locus) that has the flexibility to select several segments (i.e., putative enhancers) from the locus, assumes that the selected segments function independently, and models the readout of the locus as a weighted sum of the readouts of the selected segments.

In the present work, having noted similar failures of the GEMSTAT model on fused constructs, we asked if a locus-level model such as GEMSTAT-GL might be sufficient to model this data set. Note that applying the GEMSTAT-GL model to these data amounts to assuming that a fused construct may function more like a gene locus, with independently functioning segments (enhancers) within, even though the construct's length is typical of a single enhancer. We optimized parameters of the GEMSTAT-GL model, under a highly constrained setting (as adopted in the original work (Samee and Sinha 2014); see Methods) and obtained accurate fits to the readouts of all constructs in the data set (Fig. 5.3). In particular, GEMSTAT-GL was successful in capturing both the inter-stripe gaps (as observed in constructs Fusion C, Spacer 200, and Spacer 1000) and the broad domain of overlapping stripe expressions (as observed in constructs Fusion A, B, and D). Interestingly, the model always selected one contributing regulatory segment from each of the two constituent enhancers in a fused construct. That it did not select multiple segments from any of the two constituent enhancers is consistent with the literature: prior experimental attempts to identify smaller functional segments within these constituent enhancers showed loss of function (Fujioka, Emi-Sarker et al. 1999). The GEMSTAT-GL model never selected a segment that straddled across the boundary

of the fused constructs, even though its optimization phase had the flexibility to do so. This raises the intriguing possibility of yet unidentified mechanisms that maintain the constituent enhancers' independence although they were fused without any spacer. Overall, the most significant conclusion from this exercise is that the regulatory function of each fused construct can be explained with a model that regards it as two independently acting enhancers and cannot be explained by regarding the construct as a single enhancer, even though the construct is very much like a single enhancer in its length and binding site content.

5.2.3 A model where repressors act only over short ranges can explain readouts of constructs with spacer sequences but fails on the other constructs

The above modeling exercise points to the existence of at least two independent regulatory segments within each fused construct. However, it does not offer a mechanistic explanation of their separate existence, when sequences similar to the fused construct are known to function as single enhancers. Previous studies have hypothesized that repressor TFs, by working over short-ranges, may confer independence to segments in the genomic DNA. In particular, it has been hypothesized that certain TFs upon binding to their cognate sites inhibit the binding of activator TFs to neighboring sites, typically to those located within 150 bps, and thereby repress the expression of the target gene. This short-range repression mechanism (SRR) can thus partition a given sequence into clusters of TF binding sites, and has been hypothesized to confer independence to enhancers in a locus (Small, Arnosti et al. 1993, Gray and Levine 1996, Kim, Martinez et al. 2013). We therefore considered the SRR mechanism as a potential explanation for the existence of independent segments within fused constructs. It should be noted however that two clusters (of sites) are considered independent under the SRR mechanism when they are sufficiently far apart so that repressors bound to one cluster do not interact with those in the other cluster. Closely located clusters may interfere in each other's regulatory effect as has been shown in (Small, Arnosti et al. 1993, Kim, Martinez et al. 2013). We should note here that, in our previous work with the locus-level model we found the SRR mechanism to be insufficient to model a gene's expression pattern as the readout of its locus. A different implementation of the SRR model (Kim, Martinez et al. 2013) was used to explain the expression pattern of the eve gene from its locus, but was unable to find a single parameterization of the model that could fit all seven stripes. Hence, it was not clear a priori if the SRR mechanism would prove to be an acceptable explanation for the independent action of the two enhancers in each fused construct.

To pursue the above line of investigation, we first noted that the two fused constructs having spacer sequences in our dataset show readouts with four clear stripes, as might be expected if the two constituent enhancers act independently. (Each is known to drive two stripes of expression by itself.) While stripe formation in the readouts of the other four fused constructs is disrupted, their overall expression domain still appears to be the same as the combined domain of the four eve stripes. We asked, in light of considerations described above, whether the SRR mechanism may explain: (a) the proper formation of four stripes in the former two cases by conferring independence to the two fused enhancers due to spacers between them, and (b) the disrupted stripe formation in the latter four cases (no spacer

between fused enhancers) by revealing interference in regulatory effects at the junctions of the fused constructs.

For each fused construct in our dataset we fit its readout using the GEMSTAT model in the SRR mode (GEMSTAT-SRR), allowing the model the same flexibilities in training as were allowed to the baseline GEMSTAT model (see Methods). We found that for the constructs with spacers the model's accuracy was comparable to that of the locus-level GEMSTAT-GL model (Fig. 5.4). However, for those where the two enhancers were fused adjacent to each other, the model's fits were not as accurate as GEMSTAT-GL, though improved over the baseline GEMSTAT model. These observations imply that the SRR model is not mechanistically rich enough to explain the readouts in our current dataset. In particular, for none of the fused constructs, we could find a satisfactory fit so that we could attribute the disrupted stripe formation to interference at the junctions. We therefore favor the GEMSTAT-GL model's findings as the explanation to the readouts of our fused constructs: linear aggregation of independent regulatory segments' readouts give rise to the observed expression patterns.

5.3 Methods

The models GEMSTAT and GEMSTAT-GL have been described in Chapters 2 and 4. Enhancer sequences and TF motifs were collected from the same sources as used in our work on GEMSTAT-GL described in Chapter 4. There are several motifs available for each TF. We therefore needed to select the motif and the threshold on the LLR-score (see Chapter 2) for each TF. To this end, we first listed all the footprinted binding sites in the enhancers for *eve* from the Redfly database (Gallo, Gerrard et al. 2011). For each TF, we selected the motif and the threshold that could identify all the footprinted sites along with the minimum number of additional weak sites. The expression profiles were collected from (Lydiard-Martin et al., manuscript in preparation). The parameter space exploration and subsequent pipeline for choosing 1000 best models is adopted from our ensemble construction method described in Chapter 3. The constrained optimization method for parameter optimization was adopted from our work on GEMSTAT-GL described in Chapter 4.

5.4 Discussion

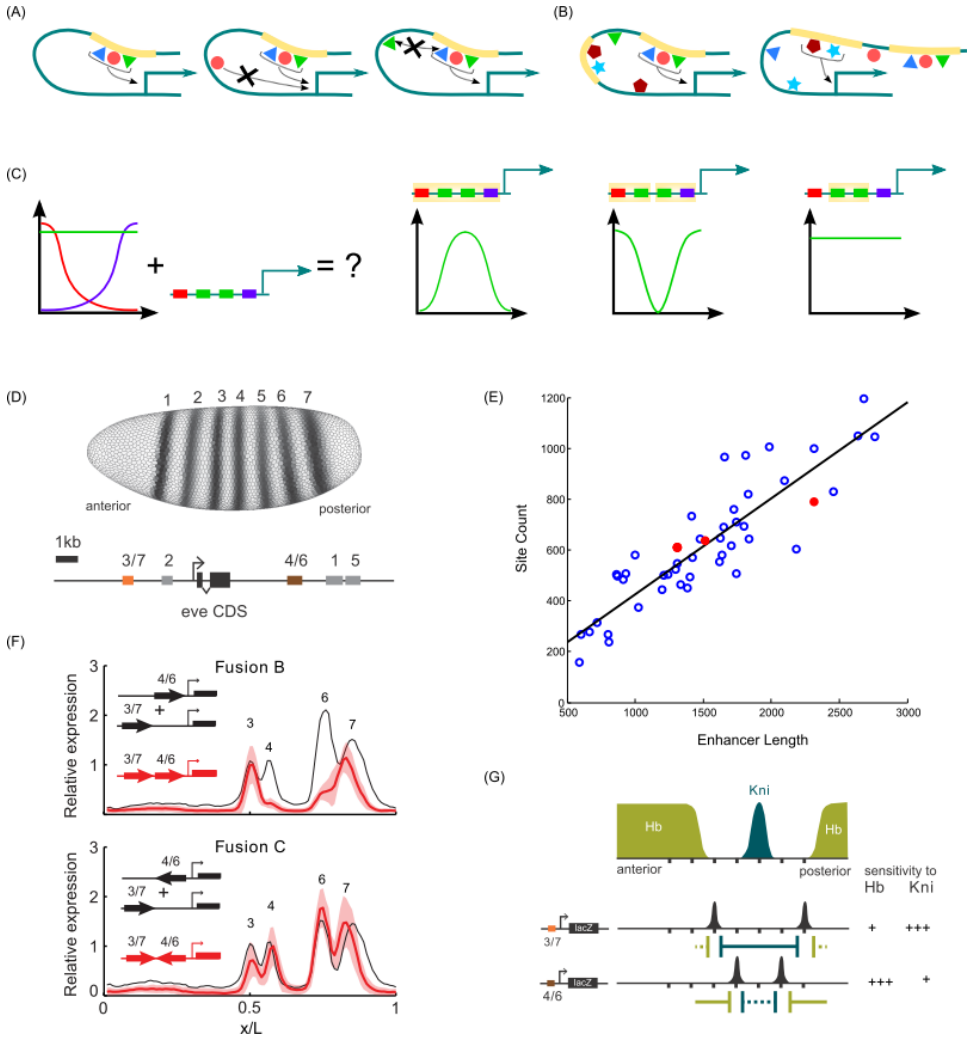
The classic definition of an enhancer is empirical: it is a piece of DNA that can direct expression of a target gene independent of orientation and distance from the promoter (Moreau, Hen et al. 1981, Banerji, Olson et al. 1983, Gillies, Morrison et al. 1983). This definition reflects the first experiments used to identify enhancers and their major properties, namely dense clustering of TF binding sites, high TF occupancy, specialized histone marks, and location within open chromatin structure (reviewed in (Bulger and Groudine 2010)). However, we are far from understanding the details of the *cis*-regulatory process mechanistically; this requires us to abstract the underlying processes appropriately because different mechanisms may be captured by different functional forms. If we abstract inappropriately, we may not capture our experimental data accurately or we may be unable to discern underlying mechanistic principles. *A priori* we can define an enhancer as a functional unit whose output depends on short-range interactions; longer range interactions between enhancers can then be modeled separately. This definition of an enhancer reflects its role in information-processing—it is a mathematical description of

how it interprets TF concentrations to produce a specific expression pattern. Our fusions demonstrate the utility of this definition. The fusions look like a single enhancer in terms of average length and overall binding site content and behave as a typical single "active" fragment in reporter assays. However, they are best modeled as two separate functional units whose output can be aggregated as a weighted sum to produce the total expression pattern of the reporter.

What molecular mechanisms might define the length of a functional segment within a given sequence? Local interactions between TFs are well-documented and could provide a length constraint on TFs that can work together. In support of this type of underlying mechanism, the "enhanceosome" is a 55 bp enhancer that works by assembling a large protein complex in a highly cooperative fashion (Panne 2008). The precise placement of TF binding sites within the enhancer facilitates assembly of the complex. Altering the position and affinity of the binding sites disrupts enhanceosome function, and as a result, the sequence of this enhancer is highly conserved. Though developmental enhancers are typically much longer (200 bps to 2 kbps or larger), they could be comprised of small cooperative units defined by local interactions, as proposed in the billboard model (Arnosti and Kulkarni 2005). There could also be length-restrictions on functional segments independent of local TF interactions. For example, chromatin state or nucleosome placement could define a region that is interpreted as a single functional unit. DNAse accessibility clearly helps to define which TF binding sites are occupied (Kaplan, Li et al. 2011). However, currently available assays for accessibility *in vivo* average the measurements for these features across multiple cells, making it difficult to discern whether they could also help to define functional segments at a fine scale. Another possible source of a length constraint could be the physical interaction with co-factors or the promoter. Presumably, a constrained amount of bound DNA is involved in these interactions. This would result in a length constraint on functional segments that is dependent on how much sequence is sampled during each interaction event, whether the same piece is always sampled, and the dynamics of the interaction. Length constraints that do not depend on local cooperative interactions would likely have different evolutionary properties. TF binding sites would be free to rearrange within the functional segment and there would be smooth rather than abrupt changes in activity due to mutations. We propose that the two particular enhancers in this study are composed of functional segments, larger than individual TF binding sites, but smaller than the component enhancer. Each of these segments can direct expression in the proper position but at a lower level than the entire enhancer. Our computational analysis finds sub-enhancer length fragments capable of driving expression in the proper position, but to varying levels. These fragments are placed together asymmetrically within the annotated enhancers. The asymmetry may explain the orientation dependence of expression that we observe. A clear future direction is to test the expression driven by these predicted fragments located within the annotated enhancers. A technical challenge will be detecting potentially low-levels of expression, though this should be aided by quantitative imaging and improvements in fluorescent staining techniques.

5.5 Figures

Figure 5.1: Overview of different enhancer-readout mechanisms and the dataset



(A) Assumptions of independence underlying the enhancer model. (B) Assumption of exclusively functioning enhancers underlying the locus model. (C) Different possible scenarios due to different delineations of independent segments in a sequence. (D) The *Drosophila* embryo and the enhancers and the expression domain of *eve*. (E) Plot showing site count vs. enhancer length for the 44 enhancers used by Segal et al. (Segal, Raveh-Sadka et al. 2008) (blue) and the constructs in the current study (red). (F) Expression readout of two example constructs. (G) The underlying mechanism of how, due to differences in affinities of the HB and KNI sites, the two enhancers drive expression in different domains.

Figure 5.2: GEMSTAT model on the data set

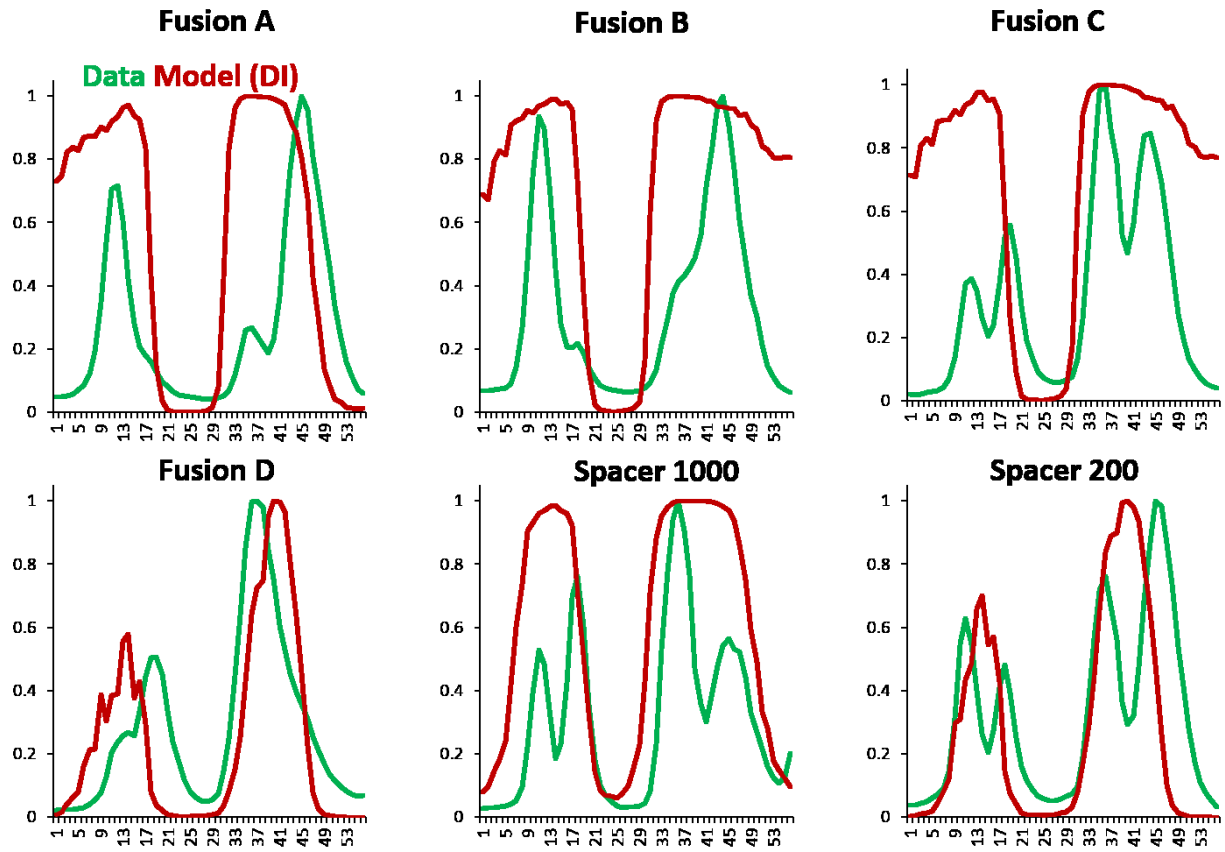


Figure 5.3: GEMSTAT-GL on the data set

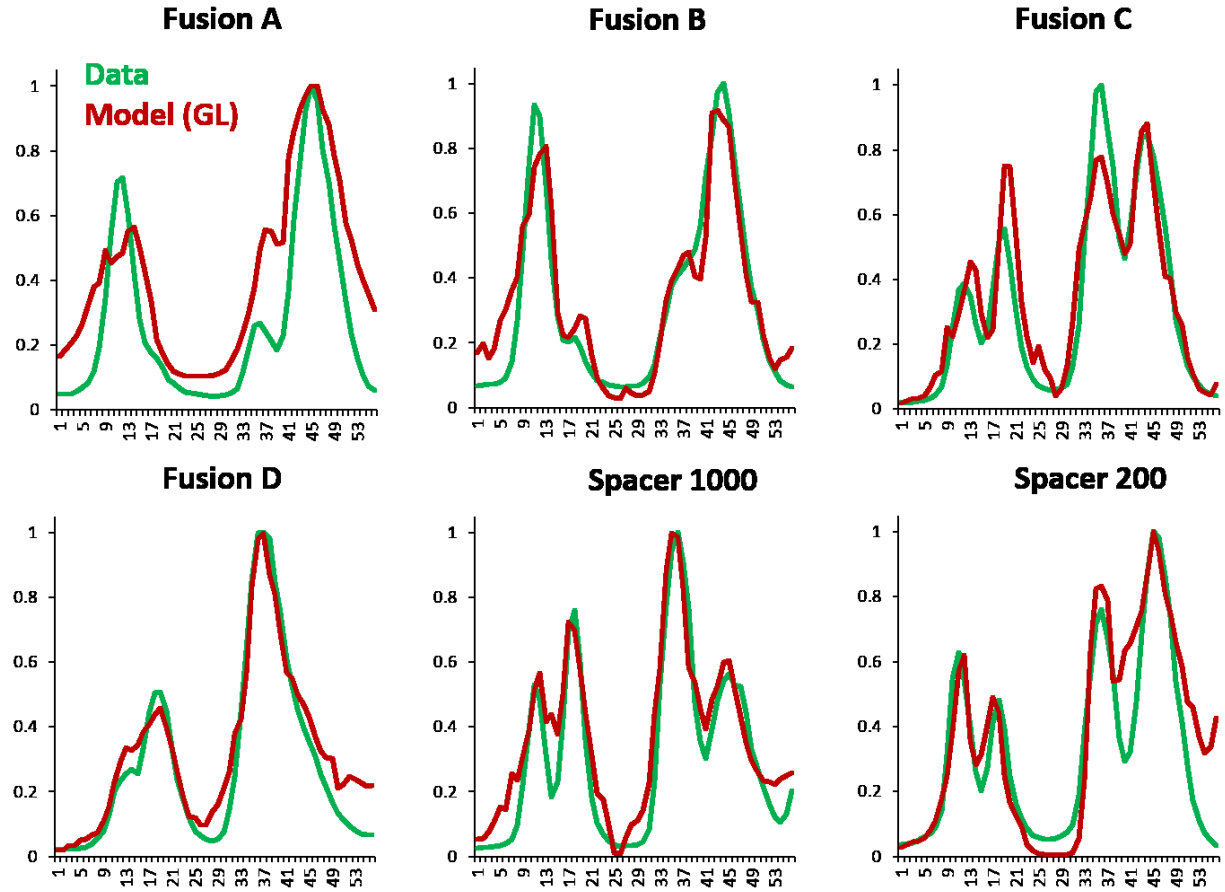
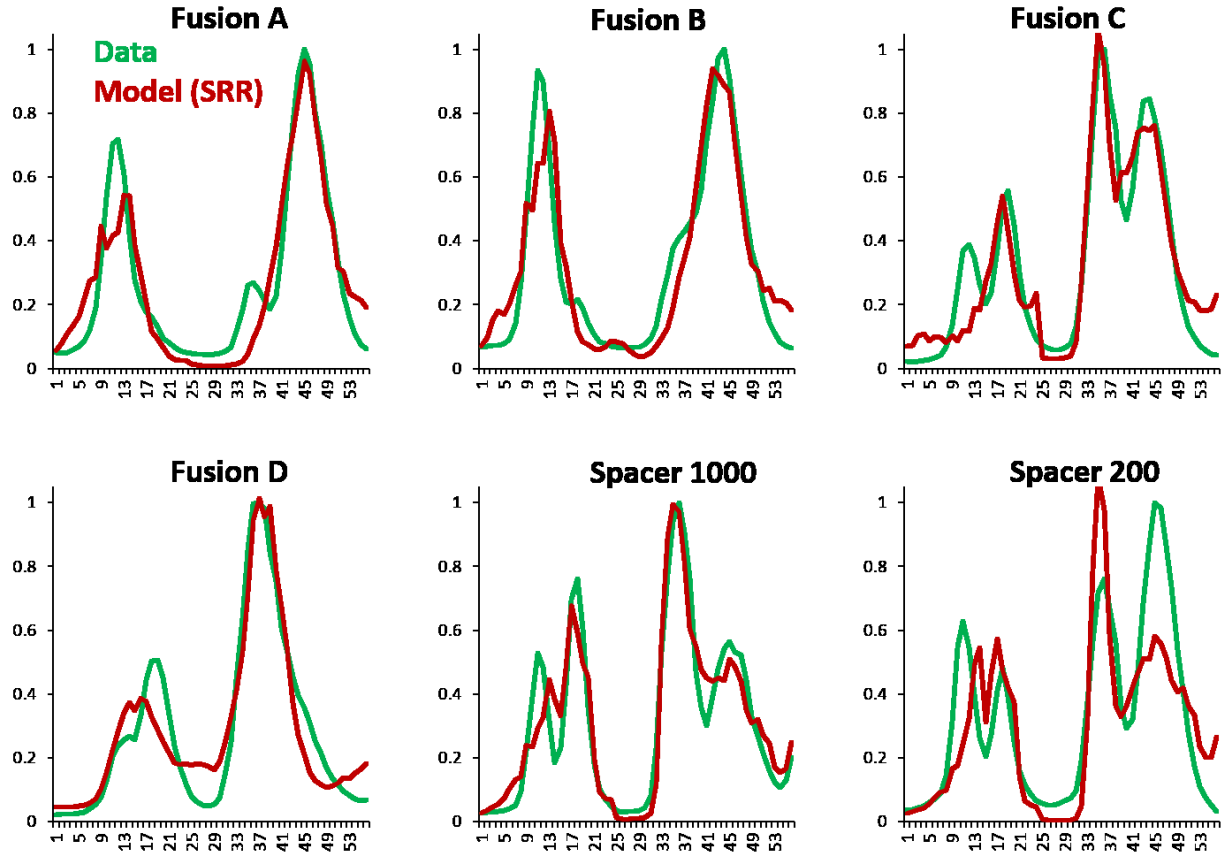


Figure 5.4: GEMSTAT-SRR on the data set



Chapter 6
Quantitative Modeling of Gene Expression in the *Drosophila* Imaginal Wing Disc

6.1 Introduction

Mathematical modeling of organ growth is as classical a topic as that of modeling embryonic pattern formation (Zhang, Alber et al. 2013). However, state of the art models of organ development focus only on the morphology of organs and do not elicit molecular explanations at the level of regulatory sequences (Yin, Xiao et al. 2013, Zhang, Alber et al. 2013). Given that the ultimate goal of developmental investigation is to understand the logic of regulatory control as encoded in the genomic DNA, i.e., to understand the *cis*-regulatory network of genes involved in patterning and growth, it is therefore important to include enhancer sequences in models of organ growth. In this work, we take the first step toward this goal by applying a state of the art sequence-to-expression model on a network of genes that are responsible for the primary patterning of the *Drosophila* wing imaginal disc and formation of the second vein (known as L2) therein (Fig. 6.1-A).

The *Drosophila* wing is a widely adopted model system in studies of organ growth, scaling, and patterning (Affolter and Basler 2007, Wartlick, Mumcu et al. 2011). Many genes involved in wing development have been identified – with concomitant hypotheses of how they work – yet no study to date has attempted to build a sequence-level understanding of how these genes are expressed. The "WingX" project – the most elaborate attempt undertaken to date for understanding wing development – plans to elicit systems-level, broad pictures of *Drosophila* wing development, but does not address the problem of charting the regulatory sequences of this system or deciphering how they control the related genes. Therefore, the comprehensiveness of existing knowledge about *cis*-regulation of the wing developmental genes has never been assessed.

We apply here the GEMSTAT model on the *cis*-regulatory network comprising five genes, namely *brinker* (*brk*), *optomotor-blind* (*omb*), *spalt* (*sal*), *daughters against dpp* (*dad*), and *knirps* (*kni*), for the primary patterning of the *Drosophila* wing imaginal disc. Our objective here is to complement the known experimental observations on these genes and quantitatively assess various hypotheses of how their expression is regulated. The models developed here will offer several additional benefits. First, upstream regulatory events that set the primary morphogen of wing development (i.e., Decapentaplegic, Dpp) are being actively studied by several other groups (Nahmad and Stathopoulos 2009, Parker, White et al. 2011). Our study, combined together with these ongoing efforts, will enable us to obtain an understanding of a larger system of genes beyond this primary morphogen. Secondly, the gene regulatory network in wing disc patterning is nearly identical to that in haltere patterning except for the expression of the gene *ultrabithorax* (Weatherbee, Halder et al. 1998). Our study thus holds the promise of explaining data from two developing organs. Furthermore, we have confirmed that the assumed structure of the gene regulatory network is reflected in its *cis*-regulatory sequences by validating the model predictions against

various perturbation data. Finally, utilizing the models, we have offered novel insights about the system of genes functioning in wing development.

6.2 Results

6.2.1 A sequence-to-expression model of the genes involved in patterning the anterior compartment of *Drosophila* wing imaginal disc

In this work, we have modeled the gene regulatory network involved in early patterning of the wing pouch in *Drosophila* imaginal wing disc. The imaginal wing disc in *Drosophila* refers to a cluster of cells in the developing embryo that are destined to form the wing blade and the notum. An oval region within the disc – called the "pouch" – corresponds to the blade; the region can be divided into four quadrants based on the horizontal (anterior-posterior, A/P) and vertical (dorso-ventral, D/V) axis of the pouch. Gene expression and developmental morphology are symmetric with respect to the A/P, but not to the D/V, axis. However, the regulatory roles played by each gene remain the same throughout the pouch. We chose here to model the part of the anterior compartment where the level of *Dpp* expression falls from the maximum to the basal level at the lateral boundary of the pouch (Fig. 6.1). Within this domain of expression, *Dpp* activates three genes, expressed in nested domains that are localized near the peak level of *Dpp* expression, namely *omb*, *sal*, and *dad*. *Dpp* also activates *brk* and localizes its expression at the lateral boundary. The gene *kni* is expressed in a narrow domain between *brk* and *sal* domains. In particular, *kni* domain is completely constrained within the *sal* domain and the lateral borders of the two genes coincide. We assumed the location of the posterior border of *kni* corresponds to the location of 10% of the peak concentration of *sal*. Although specific regions within the wing pouch are under control of signaling networks, for these genes within the spatial region of our interest, it is widely known to be transcriptional control, as shown in Fig. 6.1. The regulatory edges in the figure have been reported in (Affolter and Basler 2007). For each of the genes we collected their experimentally validated enhancer and for each TF we collected their sequence specificity. We scanned the enhancers for the putative sites of these TFs and the scans supported these edges. Of note in Fig. 6.1 is the inclusion of the protein Scalloped (Sd) which is ubiquitously expressed and known to regulate several genes in the wing development by forming a complex with another protein named Vestigial (Vg). This mode of Sd-Vg complex is specific to wing development and has been shown to be important for the expression of several genes during early wing development. There was no motif for Sd-Vg complex, hence we created a motif from the experimentally tested sequences that were reported in (Guss, Nelson et al. 2001) (see Methods). The concentration profiles of the TFs were collected from images published in (Moser and Campbell 2005) and processed using the image processing software ImageJ (Rasband 1997-2014) (see Methods).

As shown in Fig. 6.2 the model fits the data accurately capturing all the notable and important features, *i.e.*, the nested domains of *omb*, *sal*, and *dad*, the precise domain of *kni*, and the narrower peak yet overall broader expression domain of *dad* in comparison to that of *sal*.

6.2.2 Model quantifies the role of each TF and the effect of its knock-down

The Role and Significance of Scalloped: The role of Scalloped (Sd) has been demonstrated to be important for several genes. However, Sd functions in developing wings as a complex with Vg. A DNA-binding motif for this complex was not well defined in the literature. We created a motif from the highly conserved sites reported in FlyReg and in (Guss, Nelson et al. 2001). The motif matches all the experimentally tested instances. Also, a match to this motif is located in the *brk* enhancer identified by Muller *et al*. (Muller, Hartmann et al. 2003). Notably, the Sd-Vg motif – despite its high information content – is present in all but two of the enhancers of *brk* that were reported in Yao *et al.* (Yao, Phin et al. 2008). Sd-Vg was also important for our model to predict a strong expression for *kni*. Overall, these observations assured us about the ability of our motif to identify functional Sd-Vg binding sites. We find that Sd-Vg sites are important for accurate modeling of expression profiles of all genes in our data set but *dad* whose enhancer does not contain any site for Sd-Vg. We find that knock-down of Sd causes abolishment of *brk* expression, while it causes weak and expanded expression for each of *omb*, *sal*, and *kni*.

The Effects of Changes in the Expression of the Dpp Morphogen: Our model captures the effect of the *Dpp* morphogen very accurately as described below (see Fig. 6-3). For *brk*, we find an expansion in its domain when *Dpp* is restricted more toward the medial axis. Under the same change, the downstream genes *omb*, *sal*, and *dad* are expressed in narrower domains. The effect on *kni* is more subtle, yet accurately captured by our model: since *kni* is not directly regulated by *Dpp*, we use the altered expression domains of the above regulators of *kni* to predict its expression and find that the *kni* domain shifts toward the direction of shift of the *sal* domain.

The Effect of Brinker Knock-Down: In this case, we find that all the genes are expanded anteriorly with the lateral borders being defined by the level of *Dpp*. As a result, we do not see a clear border of *kni* being formed.

6.2.3 Model details the mechanisms of *cis*-regulation of the genes by regulatory sequences within their loci

Mechanisms of brinker *Regulation*: Our motif scanning identified both Sd-Vg and Schnurri-Mad-Medea (SMM) sites in the *brk* locus. In particular, these sites were there in every enhancer reported in the current literature (Muller, Hartmann et al. 2003, Yao, Phin et al. 2008). Our model predictions from each of these enhancers was accurate (Fig. 6-4). We thus concluded that *brk* is activated by Sd-Vg and repressed by Dpp. We note that, the activation of *brk* was hitherto poorly understood. Ours is the first demonstration at a quantitative level, with these validation results, that Sd-Vg is a strong candidate for being considered as an activator of brk. (Fig. 6-4)

Mechanisms of spalt *Regulation*: The original study that identified the regulatory sequences of *sal* (Barrio and de Celis 2004) reported two enhancers. Our model predicts the correct expression pattern from both of these enhancers. The same study identified repressor sites of Brk in the *sal* locus by identifying several other sequences that drive expanded expression of *sal*. Our model

predicts the correct readout from these sequences as well. However, for the activator sequence in the *sal* enhancer, Barrio and de Celis could not identify the activator TF. Our model identifies Sd-Vg as the activator TF in this sequence and this also matches with the activator sequence identified by Barrio and de Celis that could drive ubiquitous *sal* expression. (Fig. 6-5)

6.2.4 Model Highlights an Important Role for Scalloped in Scaling of Gene Expression with Organ Growth

It is known that the expression of Dpp and its downstream genes scale with the size of the developing *Drosophila* wing. However, the mechanism of how the gene expression domains might scale is not clear (Ben-Zvi, Pyrowolakis et al. 2011). Recent works have shown that scaling of Dpp expression is due to the presence of Pentagone which helps in diffusion of Dpp. The downstream genes in this cascade, however, are known to be directly transcribed in a graded expression domain rather than being diffused as is the case with Dpp. More interestingly, Dpp does not only scale with tissue size but also increases in its expression level (Hamaratoglu, de Lachapelle et al. 2011) – which makes it impossible to explain the patterning of the target genes under the French-flag model (Wolpert 1969, Jaeger and Martinez-Arias 2009). We note that a dynamic increase in Sd's expression can reconcile all these confusion and present a coherent picture: an increase in the expression level of Sd in proportion with the scaled and increased expression level of Dpp can ensure the proper scaling of *brk* pattern. This in turn can ensure the proper scaling of all the downstream genes.

6.3 Methods

The motif for Sd was created by collecting the published Sd-Vg binding sites from (Guss, Nelson et al. 2001) and the highly conserved sites from FlyReg (Bergman, Carlson et al. 2005). The model fitting was done using the same pipeline as described in Chapter 2. We had one additional free parameter in the model that captured the extent beyond which Sal does not repress *kni*.

6.4 Discussion

Our work demonstrates for the first time the importance of Sd in patterning the Drosophila wing disc at a quantitative level. The high conformance of our model predictions with available data from different sources make us confident about a real role of Sd in this case. There is an increasing interest in discovering how uniformly expressed TFs may facilitate the regulation of target genes. We showed here how a uniformly expressed TFs may be important for proper scaling of the developing tissue as well. Furthermore, as has been shown with other morphogen gradients, we showed that the domains of the morphogen targets do not only follow the morphogen's spatial pattern, rather it is decided by combinatorial interaction between the morphogen, and graded and uniformly expressed TFs

6.5 Figures

Figure 6.1: Network of the modeled genes and the expression patterns thereof

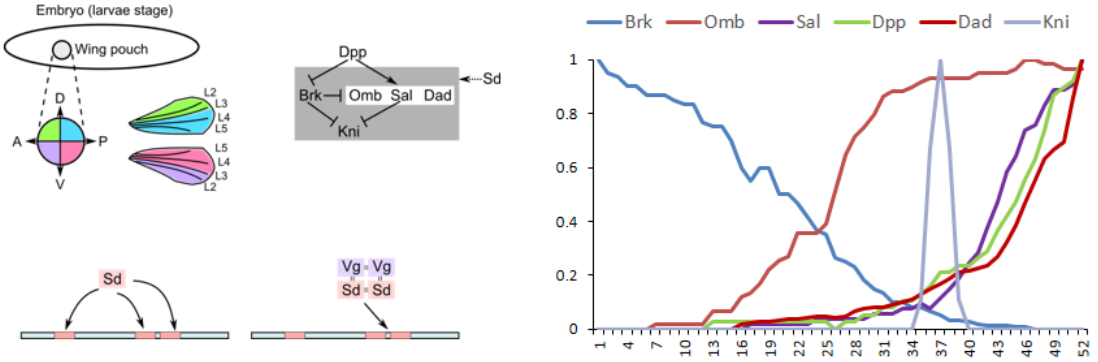


Figure 6.2: Results of model fitting on the five genes' expression pattern

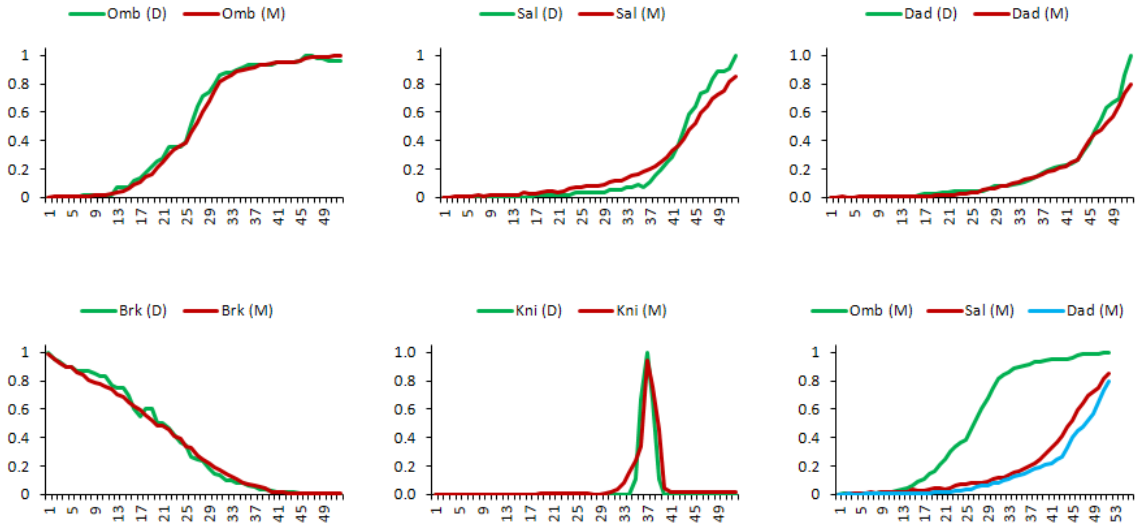


Figure 6.3: Model predictions of changing the Dpp expression domain

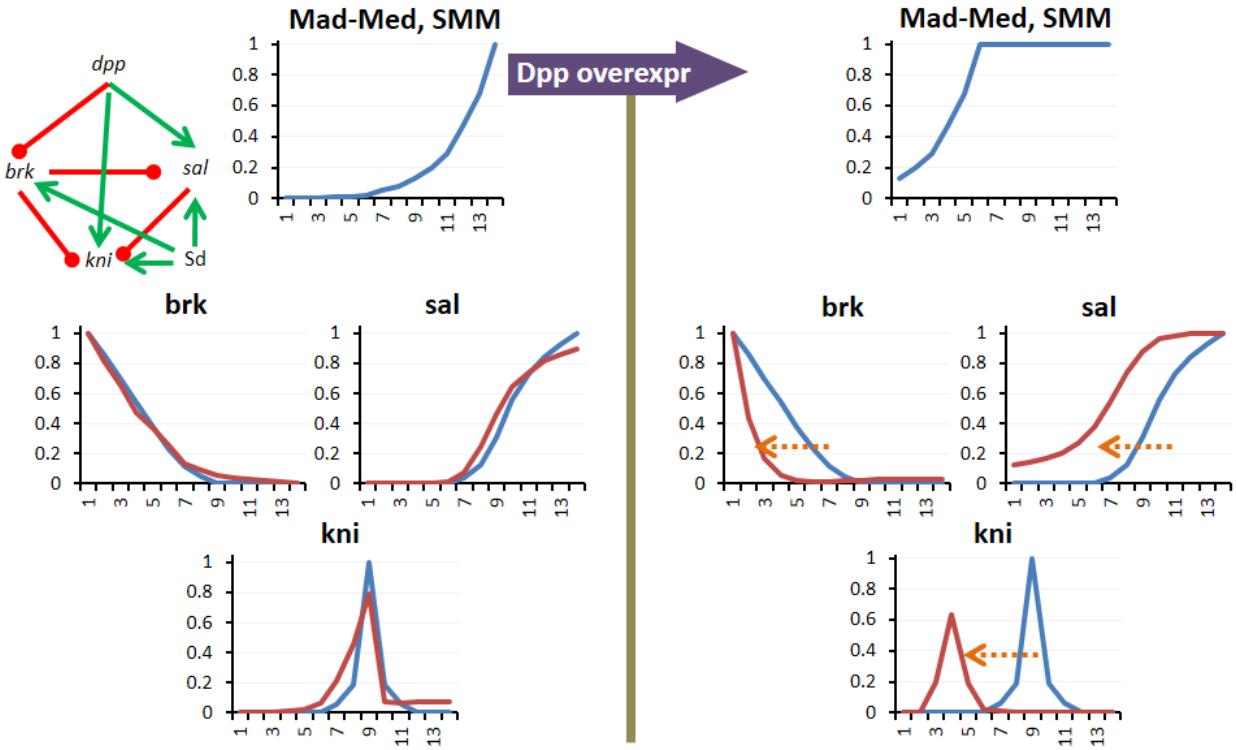
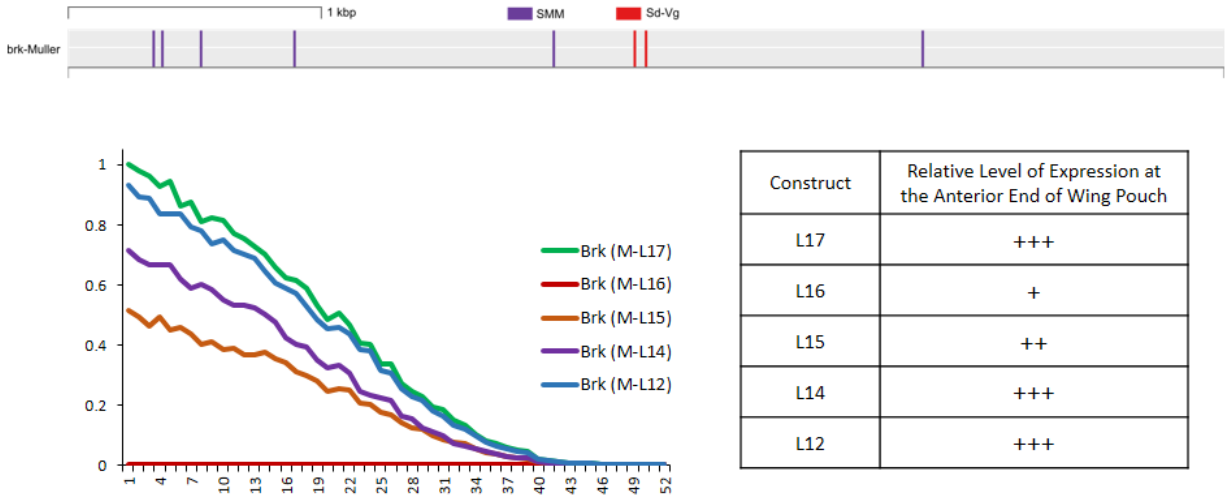
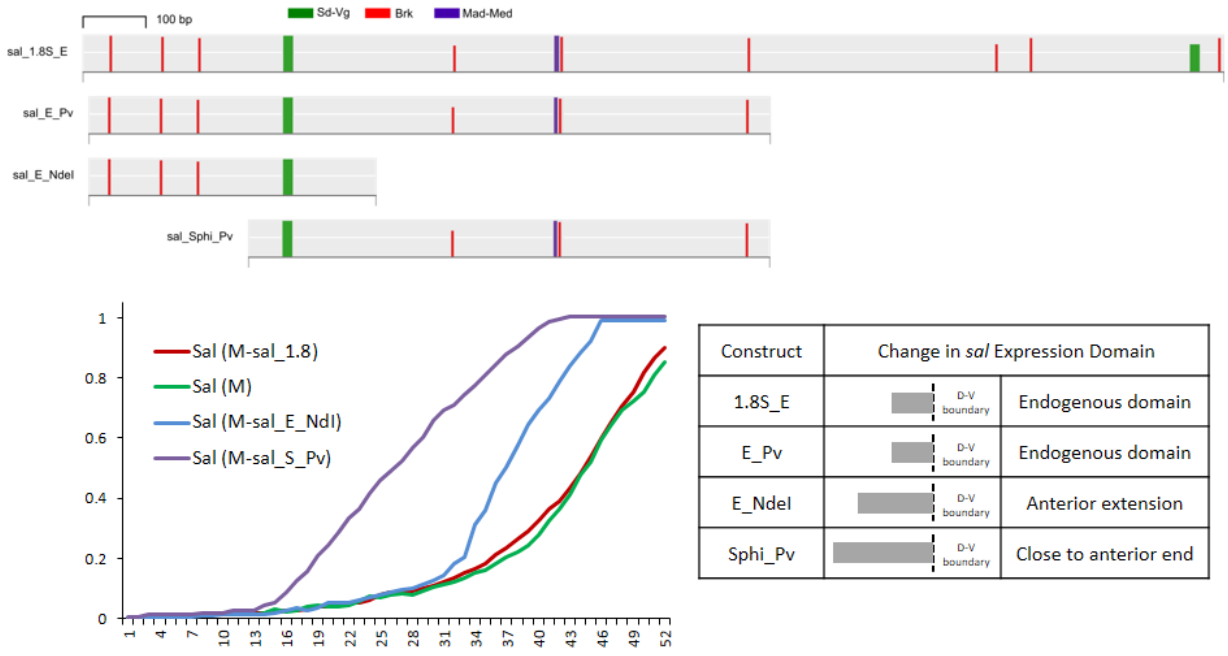


Figure 6.4: Predictions of brk expression from the known enhancers of brk



Construct	Relative Level of Expression at the Anterior End of Wing Pouch
L17	+++
L16	+
L15	++
L14	+++
L12	+++

Figure 6.5: Predictions of sal expression from the different sequences tested in (Barrio and de Celis 2004)



Construct	Change in *sal* Expression Domain	
1.8S_E	D-V boundary	Endogenous domain
E_Pv	D-V boundary	Endogenous domain
E_NdeI	D-V boundary	Anterior extension
Sphi_Pv	D-V boundary	Close to anterior end

Chapter 7
Conclusion

Understanding how DNA encodes instructions for cellular function has been a major interest in biological research ever since the onset of genome sequencing projects. Recent works highlighting the roles of mutations in non-coding DNA have put more thrust in this direction. The contribution of my thesis may be summarized as follows.

We introduce an *ensemble* based approach for modeling of transcriptional *cis*-regulation. Computational models of gene expression from *cis*-regulatory sequences are becoming increasingly complex in both their structures and the number of parameters, in trying to keep up with our growing knowledge of regulatory inputs to a gene. Models of comparable complexity commonly used in other disciplines, from signaling networks to climatology, are known to have many parameterizations that are consistent with available data. While different parameterizations may encode distinct hypotheses, the current practice in modeling transcriptional *cis*-regulation is to identify one model that best fits the data. Using the *Drosophila ind* (*intermediate neuroblast defective*) gene as our working example, we demonstrate how this practice is prone to providing an incomplete picture of reality and even lead to incorrect conclusions. We instead present an approach where we systematically explore the entire multi-dimensional parameter space and leverage information from perturbation experiments to construct an *ensemble* of models that are consistent with available information on *ind* regulation. The value of this approach, as we discuss in the here, is not only in robustly identifying models consistent with both wild-type and mutant data, but also in pinpointing further experiments that will eliminate current gaps in our understanding of the gene's regulation. The methodological significance of our work is thus as an example of how models in this realm should be built and analyzed in future.

In computing models that unify the current knowledge of *ind* regulation, we believe we have made an important contribution of interest to the developmental biology community and more broadly to scientists working to decipher the "*cis*-regulatory code". The *ind* gene is an early developmental gene whose transcription depends on both extra-cellular signaling (MAPK) and the combinatorial action of an assortment of transcription factors, both activators and repressors. Our models justify a mechanism – proposed based on similar results from studies on cultured human cells – of how the MAPK signaling may affect DNA binding affinity of the transcription factor Capicua. To our knowledge, this is the first quantitative work that uses sequence-level modeling to demonstrate the involvement of signaling pathways in regulation of gene transcription.

Another important biological contribution of this work is to show how Zelda, a ubiquitously expressed 'pioneer' transcription factor, may regulate transcription upon binding to the DNA. We report experimental validation of our prediction about the quantitative effect of inhibiting Zelda binding to the *ind* enhancer. This result, based on a combination of modeling and experimental work, is especially interesting in light of recent experimental studies demonstrating a role for Zelda in remodeling the chromatin, affecting DNA accessibility, and ultimately influencing gene expression in the early *Drosophila* embryo.

We also establish the importance of weak-affinity (non-consensus) binding sites of the transcription factor Dorsal in activating *ind* and show that while our model-based predictions are in agreement with *cis*-perturbation experiments conducted previously, the conclusion drawn from those experiments regarding the role of Dorsal may have been incorrect, in part due to a disregard of the contribution of weak-affinity sites. We believe this will be appreciated as an important example aligned with the recent findings about evolutionary and functional significance of weak-affinity sites.

We present a computational framework for interpreting the sequence of a gene's intergenic region and modeling the gene's expression level in a cell type, given the concentrations of relevant transcription factors in that cell type. The quantitative model builds upon our previous work on statistical thermodynamics-based modeling of enhancer function, which was published two years ago (6). The new model relies upon those thermodynamics-based models to predict the readout of individual enhancers, and uses a weighted summation of those readouts to predict the regulatory output of the entire locus. We demonstrate the utility of the model in predicting the complex multi-stripe expression patterns of several genes in early embryonic development in *Drosophila melanogaster.* The model automatically discovers segments within the locus that contribute to gene expression. Segments that are not selected are in some cases predicted to produce a pattern that is irreconcilable with the gene's overall expression pattern, and must be "shut down" by some mechanism. Indeed, these non-selected segments are observed to be in inaccessible regions of the locus. In light of various in silico experiments conducted using the new model, we argue that the full complement of binding sites spread out over the locus acts together in ways that are different from the combinatorial action of sites within any one enhancer.

We expect this new work to push the boundaries of quantitative modeling of gene expression, and spur further work on sequence-to-expression models that operate without prior knowledge of enhancer locations. This will be the natural next step in the community's attempts to lay out the cis-regulatory code in quantitative terms. We also anticipate our computational framework to play a significant role in synthetic biology, which needs to precisely quantify input-output relationships of regulatory sequences, in studies of regulatory evolution, and in the common bioinformatics task of discovering novel regulatory elements.

Some future extensions of this modeling framework are as follows. First, the framework may be extended to understand the basis of tissue- and stage-specific gene expression. Information on co-factor recruitment along with chromatin remodeling, three-dimensional configuration of the DNA, and various histone modification marks may aid to this end. However, the challenge will be to understand how these new factors may be integrated in a biophysical model, or whether we can use them more phenomenologically as we did to incorporate MAPK's action on Cic. It is also worth investigating whether and to what extent the above functional genomics data are generalizable to arbitrary contexts of interest. A second extension of this system is in the realm of dynamic studies of gene expression. Clearly there is a thrust in this direction to understand variation in gene expression due both to intrinsic and extrinsic noise through dynamic models. None of these models, however, consider the effect of sequence – which we have shown in Chapter 3 as an important factor. Recent single cell genomics data, coupled with our mechanistic framework will be valuable in the future. A major application of this system will be in identifying causal variants among the significant SNPs and variants identified through statistical

overrepresentation. The ultimate goal of GWAS studies can only be realized when one can explain functional associations – for which a sequence to expression framework like ours would be valuable.

References

Affolter, M. and K. Basler (2007). "The Decapentaplegic morphogen gradient: from pattern formation to growth regulation." Nat Rev Genet **8**(9): 663-674.

Agarwal, S., L. Jongwoo, L. Zelnik-Manor, P. Perona, D. Kriegman and S. Belongie (2005). Beyond pairwise clustering. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.

Ajuria, L., C. Nieva, C. Winkler, D. Kuo, N. Samper, M. J. Andreu, A. Helman, S. Gonzalez-Crespo, Z. Paroush, A. J. Courey and G. Jimenez (2011). "Capicua DNA-binding sites are general response elements for RTK signaling in Drosophila." Development **138**(5): 915-924.

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

Arnold, C. D., D. Gerlach, C. Stelzer, L. M. Boryn, M. Rath and A. Stark (2013). "Genome-wide quantitative enhancer activity maps identified by STARR-seq." Science **339**(6123): 1074-1077.

Arnosti, D. N. (2003). "Analysis and function of transcriptional regulatory elements: insights from Drosophila." Annu Rev Entomol **48**: 579-602.

Arnosti, D. N., S. Barolo, M. Levine and S. Small (1996). "The eve stripe 2 enhancer employs multiple modes of transcriptional synergy." Development **122**(1): 205-214.

Arnosti, D. N. and M. M. Kulkarni (2005). "Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?" J Cell Biochem **94**(5): 890-898.

Ay, A. and D. N. Arnosti (2011). "Mathematical modeling of gene expression: a guide for the perplexed biologist." Crit Rev Biochem Mol Biol **46**(2): 137-151.

Banerji, J., L. Olson and W. Schaffner (1983). "A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes." Cell **33**(3): 729-740.

Barolo, S. (2012). "Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy." Bioessays **34**(2): 135-141.

Barolo, S. and M. Levine (1997). "hairy mediates dominant repression in the Drosophila embryo." EMBO J **16**(10): 2883-2891.

Barolo, S. and J. W. Posakony (2002). "Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling." Genes Dev **16**(10): 1167-1181.

Barrio, R. and J. F. de Celis (2004). "Regulation of spalt expression in the Drosophila wing blade in response to the Decapentaplegic signaling pathway." Proc Natl Acad Sci U S A **101**(16): 6021-6026.

Ben-Zvi, D., G. Pyrowolakis, N. Barkai and B. Z. Shilo (2011). "Expansion-repression mechanism for scaling the Dpp activation gradient in Drosophila wing imaginal discs." Curr Biol **21**(16): 1391-1396.

Berg, O. G. and P. H. von Hippel (1987). "Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters." J Mol Biol **193**(4): 723-750.

Bergman, C. M., J. W. Carlson and S. E. Celniker (2005). "Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster." Bioinformatics **21**(8): 1747-1749.

Berman, B. P., Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin and M. B. Eisen (2002). "Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome." Proc Natl Acad Sci U S A **99**(2): 757-762.

Berman, B. P., B. D. Pfeiffer, T. R. Laverty, S. L. Salzberg, G. M. Rubin, M. B. Eisen and S. E. Celniker (2004). "Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura." Genome Biol **5**(9): R61.

Bintu, L., N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev and R. Phillips (2005). "Transcriptional regulation by the numbers: models." Curr Opin Genet Dev **15**(2): 116-124.

Blackwood, E. M. and J. T. Kadonaga (1998). "Going the distance: a current view of enhancer action." Science **281**(5373): 60-63.

Brown, J. L. and C. Wu (1993). "Repression of Drosophila pair-rule segmentation genes by ectopic expression of tramtrack." Development **117**(1): 45-58.

Buchler, N. E., U. Gerland and T. Hwa (2003). "On schemes of combinatorial transcription logic." Proc Natl Acad Sci U S A **100**(9): 5136-5141.

Bulger, M. and M. Groudine (2010). "Enhancers: the abundance and function of regulatory sequences beyond promoters." Dev Biol **339**(2): 250-257.

Cheng, Q., M. Kazemian, H. Pham, C. Blatti, S. E. Celniker, S. A. Wolfe, M. H. Brodsky and S. Sinha (2013). "Computational Identification of Diverse Mechanisms Underlying Transcription Factor-DNA Occupancy." PLoS Genet **9**(8): e1003571.

Chopra, V. S. and M. Levine (2009). "Combinatorial patterning mechanisms in the Drosophila embryo." Brief Funct Genomic Proteomic **8**(4): 243-249.

Cornell, R. A. and T. V. Ohlen (2000). "Vnd/nkx, ind/gsh, and msh/msx: conserved regulators of dorsoventral neural patterning?" Curr Opin Neurobiol **10**(1): 63-71.

Courey, A. J. (2008). Mechanisms in transcriptional regulation. Malden, MA ; Oxford, Blackwell Pub.

Cowden, J. and M. Levine (2003). "Ventral dominance governs sequential patterns of gene expression across the dorsal-ventral axis of the neuroectoderm in the Drosophila embryo." Dev Biol **262**(2): 335-349.

Davidson, E. H. (2006). The regulatory genome : gene regulatory networks in development and evolution. Burlington, MA ; San Diego, Academic.

Davidson, E. H. (2010). "Emerging properties of animal gene regulatory networks." Nature **468**(7326): 911-920.

DePamphilis, M. L. (2002). Gene expression at the beginning of animal development. Amsterdam ; New York, Elsevier.

Dissanayake, K., R. Toth, J. Blakey, O. Olsson, D. G. Campbell, A. R. Prescott and C. MacKintosh (2011). "ERK/p90(RSK)/14-3-3 signalling has an impact on expression of PEA3 Ets transcription factors via the transcriptional repressor capicua." Biochem J **433**(3): 515-525.

Donaldson, I. J., M. Chapman and B. Gottgens (2005). "TFBScluster: a resource for the characterization of transcriptional regulatory networks." Bioinformatics **21**(13): 3058-3059.

Dresch, J., M. Thompson, D. Arnosti and C. Chiu (2013). "Two-Layer Mathematical Modeling of Gene Expression: Incorporating DNA-Level Information and System Dynamics." SIAM Journal on Applied Mathematics **73**(2): 804-826.

Dresch, J. M., X. Liu, D. N. Arnosti and A. Ay (2010). "Thermodynamic modeling of transcription: sensitivity analysis differentiates biological mechanism from mathematical model-induced effects." BMC Syst Biol **4**: 142.

Epstein, D. J. (2009). "Cis-regulatory mutations in human disease." Brief Funct Genomic Proteomic **8**(4): 310-316.

Ernst, J. and M. Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nat Biotechnol **28**(8): 817-825.

Ernst, J., P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis and B. E. Bernstein (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types." Nature **473**(7345): 43-49.

Fakhouri, W. D., A. Ay, R. Sayal, J. Dresch, E. Dayringer and D. N. Arnosti (2010). "Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo." Mol Syst Biol **6**: 341.

Foo, S. M., Y. Sun, B. Lim, R. Ziukaite, K. O'Brien, C. Y. Nien, N. Kirov, S. Y. Shvartsman and C. A. Rushlow (2014). "Zelda potentiates morphogen activity by increasing chromatin accessibility." Curr Biol **24**(12): 1341-1346.

Frankel, N., G. K. Davis, D. Vargas, S. Wang, F. Payre and D. L. Stern (2010). "Phenotypic robustness conferred by apparently redundant transcriptional enhancers." Nature **466**(7305): 490-493.

Frith, M. C., M. C. Li and Z. Weng (2003). "Cluster-Buster: Finding dense clusters of motifs in DNA sequences." Nucleic Acids Res **31**(13): 3666-3668.

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

Gallo, S. M., D. T. Gerrard, D. Miner, M. Simich, B. Des Soye, C. M. Bergman and M. S. Halfon (2011). "REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila." Nucleic Acids Res **39**(Database issue): D118-123.

Garcia, M. and A. Stathopoulos (2011). "Lateral gene expression in Drosophila early embryos is supported by Grainyhead-mediated activation and tiers of dorsally-localized repression." PLoS One **6**(12): e29172.

Gertz, J., E. D. Siggia and B. A. Cohen (2009). "Analysis of combinatorial cis-regulation in synthetic and genomic promoters." Nature **457**(7226): 215-218.

Gillies, S. D., S. L. Morrison, V. T. Oi and S. Tonegawa (1983). "A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene." Cell **33**(3): 717-728.

Gonzalez, F., D. Duboule and F. Spitz (2007). "Transgenic analysis of Hoxd gene regulation during digit development." Dev Biol **306**(2): 847-859.

Granek, J. A. and N. D. Clarke (2005). "Explicit equilibrium modeling of transcription-factor binding and gene regulation." Genome Biol **6**(10): R87.

Grimm, O., V. Sanchez Zini, Y. Kim, J. Casanova, S. Y. Shvartsman and E. Wieschaus (2012). "Torso RTK controls Capicua degradation by changing its subcellular localization." Development **139**(21): 3962-3968.

Guss, K. A., M. Benson, N. Gubitosi, K. Brondell, K. Broadie and J. B. Skeath (2013). "Expression and function of scalloped during Drosophila development." Dev Dyn **242**(7): 874-885.

Guss, K. A., C. E. Nelson, A. Hudson, M. E. Kraus and S. B. Carroll (2001). "Control of a genetic regulatory network by a selector gene." Science **292**(5519): 1164-1167.

Gutenkunst, R. N., J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers and J. P. Sethna (2007). "Universally sloppy parameter sensitivities in systems biology models." PLoS Comput Biol **3**(10): 1871-1878.

Halfon, M. S., Y. Grad, G. M. Church and A. M. Michelson (2002). "Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model." Genome Res **12**(7): 1019-1028.

Hamaratoglu, F., A. M. de Lachapelle, G. Pyrowolakis, S. Bergmann and M. Affolter (2011). "Dpp signaling activity requires Pentagone to scale with tissue size in the growing Drosophila wing imaginal disc." PLoS Biol **9**(10): e1001182.

Harding, K., T. Hoey, R. Warrior and M. Levine (1989). "Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila." EMBO J **8**(4): 1205-1212.

Harding, K., C. Rushlow, H. J. Doyle, T. Hoey and M. Levine (1986). "Cross-regulatory interactions among pair-rule genes in Drosophila." Science **233**(4767): 953-959.

Harrison, M. M., X. Y. Li, T. Kaplan, M. R. Botchan and M. B. Eisen (2011). "Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition." PLoS Genet **7**(10): e1002266.

He, X., M. A. Samee, C. Blatti and S. Sinha (2010). "Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression." PLoS Comput Biol **6**(9).

Hong, J. W., D. A. Hendrix and M. S. Levine (2008). "Shadow enhancers as a source of evolutionary novelty." Science **321**(5894): 1314.

Hong, J. W., D. A. Hendrix, D. Papatsenko and M. S. Levine (2008). "How the Dorsal gradient works: insights from postgenome technologies." Proc Natl Acad Sci U S A **105**(51): 20072-20076.

Howard, K., P. Ingham and C. Rushlow (1988). "Region-specific alleles of the Drosophila segmentation gene hairy." Genes Dev **2**(8): 1037-1046.

Howard, K. R. and G. Struhl (1990). "Decoding positional information: regulation of the pair-rule gene hairy." Development **110**(4): 1223-1231.

Ishihara, T., S. Sato, K. Ikeda, H. Yajima and K. Kawakami (2008). "Multiple evolutionarily conserved enhancers control expression of Eya1." Dev Dyn **237**(11): 3142-3156.

Istrail, S. and E. H. Davidson (2005). "Logic functions of the genomic cis-regulatory code." Proc Natl Acad Sci U S A **102**(14): 4954-4959.

Jaeger, J. and A. Martinez-Arias (2009). "Getting the measure of positional information." PLoS Biol **7**(3): e81.

Jakobsen, J. S., M. Braun, J. Astorga, E. H. Gustafson, T. Sandmann, M. Karzynski, P. Carlsson and E. E. Furlong (2007). "Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network." Genes Dev **21**(19): 2448-2460.

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

Jiang, J., T. Hoey and M. Levine (1991). "Autoregulation of a segmentation gene in Drosophila: combinatorial interaction of the even-skipped homeo box protein with a distal enhancer element." Genes Dev **5**(2): 265-277.

Kanodia, J. S., H. L. Liang, Y. Kim, B. Lim, M. Zhan, H. Lu, C. A. Rushlow and S. Y. Shvartsman (2012). "Pattern formation by graded and uniform signals in the early Drosophila embryo." Biophys J **102**(3): 427-433.

Kaplan, N., I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom and E. Segal (2009). "The DNA-encoded nucleosome organization of a eukaryotic genome." Nature **458**(7236): 362-366.

Kaplan, T., X. Y. Li, P. J. Sabo, S. Thomas, J. A. Stamatoyannopoulos, M. D. Biggin and M. B. Eisen (2011). "Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development." PLoS Genet **7**(2): e1001290.

Kazemian, M., C. Blatti, A. Richards, M. McCutchan, N. Wakabayashi-Ito, A. S. Hammonds, S. E. Celniker, S. Kumar, S. A. Wolfe, M. H. Brodsky and S. Sinha (2010). "Quantitative analysis of the Drosophila segmentation regulatory network using pattern generating potentials." PLoS Biol **8**(8).

Kazemian, M., H. Pham, S. A. Wolfe, M. H. Brodsky and S. Sinha (2013). "Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development." Nucleic Acids Res **41**(17): 8237-8252.

Kazemian, M., Q. Zhu, M. S. Halfon and S. Sinha (2011). "Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison." Nucleic Acids Res **39**(22): 9463-9472.

Kharchenko, P. V., A. A. Alekseyenko, Y. B. Schwartz, A. Minoda, N. C. Riddle, J. Ernst, P. J. Sabo, E. Larschan, A. A. Gorchakov, T. Gu, D. Linder-Basso, A. Plachetka, G. Shanower, M. Y. Tolstorukov, L. J. Luquette, R. Xi, Y. L. Jung, R. W. Park, E. P. Bishop, T. K. Canfield, R. Sandstrom, R. E. Thurman, D. M. MacAlpine, J. A. Stamatoyannopoulos, M. Kellis, S. C. Elgin, M. I. Kuroda, V. Pirrotta, G. H. Karpen and P. J. Park (2011). "Comprehensive analysis of the chromatin landscape in Drosophila melanogaster." Nature **471**(7339): 480-485.

Kim, A. R., C. Martinez, J. Ionides, A. F. Ramos, M. Z. Ludwig, N. Ogawa, D. H. Sharp and J. Reinitz (2013). "Rearrangements of 2.5 kilobases of noncoding DNA from the Drosophila even-skipped locus define predictive rules of genomic cis-regulatory logic." PLoS Genet **9**(2): e1003243.

Kirk, P., T. Thorne and M. P. Stumpf (2013). "Model selection in systems and synthetic biology." Curr Opin Biotechnol **24**(4): 767-774.

Kirstein, M., L. Sanz, S. Quinones, J. Moscat, M. T. Diaz-Meco and J. Saus (1996). "Cross-talk between different enhancer elements during mitogenic induction of the human stromelysin-1 gene." J Biol Chem **271**(30): 18231-18236.

Kumar, S., C. Konikoff, B. Van Emden, C. Busick, K. T. Davis, S. Ji, L. W. Wu, H. Ramos, T. Brody, S. Panchanathan, J. Ye, T. L. Karr, K. Gerold, M. McCutchan and S. J. Newfeld (2011). "FlyExpress: visual mining of spatiotemporal patterns for genes and publications in Drosophila embryogenesis." Bioinformatics **27**(23): 3319-3320.

Kwasnieski, J. C., I. Mogno, C. A. Myers, J. C. Corbo and B. A. Cohen (2012). "Complex effects of nucleotide variants in a mammalian cis-regulatory element." Proc Natl Acad Sci U S A **109**(47): 19498-19503.

La Rosee, A., T. Hader, H. Taubert, R. Rivera-Pomar and H. Jackle (1997). "Mechanism and Bicoid-dependent control of hairy stripe 7 expression in the posterior region of the Drosophila embryo." EMBO J **16**(14): 4403-4411.

Levo, M. and E. Segal (2014). "In pursuit of design principles of regulatory sequences." Nat Rev Genet **15**(7): 453-468.

Liang, H. L., C. Y. Nien, H. Y. Liu, M. M. Metzstein, N. Kirov and C. Rushlow (2008). "The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila." Nature **456**(7220): 400-403.

Liberman, L. M. and A. Stathopoulos (2009). "Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence." Dev Biol **327**(2): 578-589.

Lifanov, A. P., V. J. Makeev, A. G. Nazina and D. A. Papatsenko (2003). "Homotypic regulatory clusters in Drosophila." Genome Res **13**(4): 579-588.

Lim, B., N. Samper, H. Lu, C. Rushlow, G. Jimenez and S. Y. Shvartsman (2013). "Kinetics of gene derepression by ERK signaling." Proc Natl Acad Sci U S A **110**(25): 10330-10335.

Maeda, R. K. and F. Karch (2011). "Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions." Curr Opin Genet Dev **21**(2): 187-193.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul and J. A. Stamatoyannopoulos (2012). "Systematic localization of common disease-associated variation in regulatory DNA." Science **337**(6099): 1190-1195.

McDonald, J. A., S. Holbrook, T. Isshiki, J. Weiss, C. Q. Doe and D. M. Mellerick (1998). "Dorsoventral patterning in the Drosophila central nervous system: the vnd homeobox gene specifies ventral column identity." Genes Dev **12**(22): 3603-3612.

Moles, C. G., P. Mendes and J. R. Banga (2003). "Parameter estimation in biochemical pathways: a comparison of global optimization methods." Genome Res **13**(11): 2467-2474.

Montavon, T., N. Soshnikova, B. Mascrez, E. Joye, L. Thevenet, E. Splinter, W. de Laat, F. Spitz and D. Duboule (2011). "A regulatory archipelago controls Hox genes transcription in digits." Cell **147**(5): 1132-1145.

Moreau, P., R. Hen, B. Wasylyk, R. Everett, M. P. Gaub and P. Chambon (1981). "The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants." Nucleic Acids Res **9**(22): 6047-6068.

Moser, M. and G. Campbell (2005). "Generating and interpreting the Brinker gradient in the Drosophila wing." Dev Biol **286**(2): 647-658.

Muller, B., B. Hartmann, G. Pyrowolakis, M. Affolter and K. Basler (2003). "Conversion of an extracellular Dpp/BMP morphogen gradient into an inverse transcriptional gradient." Cell **113**(2): 221-233.

Nahmad, M. and A. Stathopoulos (2009). "Dynamic interpretation of hedgehog signaling in the Drosophila wing disc." PLoS Biol **7**(9): e1000202.

Negre, N., C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller, U. Wagner, P. Kheradpour, M. L. Eaton, P. Loriaux, R. Sealfon, Z. Li, H. Ishii, R. F. Spokony, J. Chen, L. Hwang, C. Cheng, R. P. Auburn, M. B. Davis, M. Domanus, P. K. Shah, C. A. Morrison, J. Zieba, S. Suchy, L. Senderowicz, A. Victorsen, N. A. Bild, A. J. Grundstad, D. Hanley, D. M. MacAlpine, M. Mannervik, K. Venken, H. Bellen, R. White, M. Gerstein, S. Russell, R. L. Grossman, B. Ren, J. W. Posakony, M. Kellis and K. P. White (2011). "A cis-regulatory map of the Drosophila genome." Nature **471**(7339): 527-531.

Nien, C. Y., H. L. Liang, S. Butcher, Y. Sun, S. Fu, T. Gocha, N. Kirov, J. R. Manak and C. Rushlow (2011). "Temporal coordination of gene networks by Zelda in the early Drosophila embryo." PLoS Genet **7**(10): e1002339.

Ochoa-Espinosa, A. and S. Small (2006). "Developmental mechanisms and cis-regulatory codes." Curr Opin Genet Dev **16**(2): 165-170.

Panne, D. (2008). "The enhanceosome." Curr Opin Struct Biol **18**(2): 236-242.

Papatsenko, D. and M. S. Levine (2008). "Dual regulation by the Hunchback gradient in the Drosophila embryo." Proc Natl Acad Sci U S A **105**(8): 2901-2906.

Parker, D. S., M. A. White, A. I. Ramos, B. A. Cohen and S. Barolo (2011). "The cis-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity." Sci Signal **4**(176): ra38.

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

Perry, M. W., A. N. Boettiger, J. P. Bothma and M. Levine (2010). "Shadow enhancers foster robustness of Drosophila gastrulation." Curr Biol **20**(17): 1562-1567.

Perry, M. W., A. N. Boettiger and M. Levine (2011). "Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo." Proc Natl Acad Sci U S A **108**(33): 13570-13575.

Philippakis, A. A., F. S. He and M. L. Bulyk (2005). "Modulefinder: a tool for computational discovery of cis regulatory modules." Pac Symp Biocomput: 519-530.

Pisarev, A., E. Poustelnikova, M. Samsonova and J. Reinitz (2009). "FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution." Nucleic Acids Res **37**(Database issue): D560-566.

Prazak, L., M. Fujioka and J. P. Gergen (2010). "Non-additive interactions involving two distinct elements mediate sloppy-paired regulation by pair-rule transcription factors." Dev Biol **344**(2): 1048-1059.

Ptashne, M. G., A. A. F. (2002). Genes and Signals, Cold Spring Harbor Laboratory Press.

Ramos, A. I. and S. Barolo (2013). "Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution." Philos Trans R Soc Lond B Biol Sci **368**(1632): 20130018.

Rasband, W. S. (1997-2014). ImageJ (http://imagej.nih.gov/ij/). National Institutes of Health, Bethesda, Maryland, USA.

Raser, J. M. and E. K. O'Shea (2004). "Control of stochasticity in eukaryotic gene expression." Science **304**(5678): 1811-1814.

Reeves, G. T. and A. Stathopoulos (2009). "Graded dorsal and differential gene regulation in the Drosophila embryo." Cold Spring Harb Perspect Biol **1**(4): a000836.

Reinitz, J. and D. H. Sharp (1995). "Mechanism of eve stripe formation." Mech Dev **49**(1-2): 133-158.

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

Roy, S., J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White and M. Kellis (2010). "Identification of functional elements and regulatory circuits by Drosophila modENCODE." Science **330**(6012): 1787-1797.

Sagai, T., M. Hosoya, Y. Mizushina, M. Tamura and T. Shiroishi (2005). "Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb." Development **132**(4): 797-803.

Samee, M. A. and S. Sinha (2013). "Evaluating thermodynamic models of enhancer activity on cellular resolution gene expression data." Methods **62**(1): 79-90.

Samee, M. A. and S. Sinha (2014). "Quantitative modeling of a gene's expression from its intergenic sequence." PLoS Comput Biol **10**(3): e1003467.

Schneider, T. D., G. D. Stormo, L. Gold and A. Ehrenfeucht (1986). "Information content of binding sites on nucleotide sequences." J Mol Biol **188**(3): 415-431.

Schroeder, M. D., C. Greer and U. Gaul (2011). "How to make stripes: deciphering the transition from non-periodic to periodic patterns in Drosophila segmentation." Development **138**(14): 3067-3078.

Schroeder, M. D., M. Pearce, J. Fak, H. Fan, U. Unnerstall, E. Emberly, N. Rajewsky, E. D. Siggia and U. Gaul (2004). "Transcriptional control in the segmentation gene network of Drosophila." PLoS Biol **2**(9): E271.

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

Sharon, E., Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger and E. Segal (2012). "Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters." Nat Biotechnol **30**(6): 521-530.

Shea, M. A. and G. K. Ackers (1985). "The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation." J Mol Biol **181**(2): 211-230.

Sherman, M. S. and B. A. Cohen (2012). "Thermodynamic State Ensemble Models of <italic>cis</italic>-Regulation." PLoS Comput Biol **8**(3): e1002407.

Shlyueva, D., G. Stampfel and A. Stark (2014). "Transcriptional enhancers: from properties to genome-wide predictions." Nat Rev Genet **15**(4): 272-286.

Sinha, S., A. S. Adler, Y. Field, H. Y. Chang and E. Segal (2008). "Systematic functional characterization of cis-regulatory motifs in human core promoters." Genome Res **18**(3): 477-488.

Sinha, S., E. van Nimwegen and E. D. Siggia (2003). "A probabilistic method to detect regulatory modules." Bioinformatics **19 Suppl 1**: i292-301.

Small, S., D. N. Arnosti and M. Levine (1993). "Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter." Development **119**(3): 762-772.

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

Spitz, F., F. Gonzalez and D. Duboule (2003). "A global control region defines a chromosomal regulatory landscape containing the HoxD cluster." Cell **113**(3): 405-417.

Stathopoulos, A. and M. Levine (2005). "Localized repressors delineate the neurogenic ectoderm in the early Drosophila embryo." Dev Biol **280**(2): 482-493.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Stormo, G. D. and D. S. Fields (1998). "Specificity, free energy and information content in protein-DNA interactions." Trends Biochem Sci **23**(3): 109-113.

Stormo, G. D. and Y. Zhao (2010). "Determining the specificity of protein-DNA interactions." Nat Rev Genet **11**(11): 751-760.

Suleimenov, Y., A. Ay, M. A. Samee, J. M. Dresch, S. Sinha and D. N. Arnosti (2013). "Global parameter estimation for thermodynamic models of transcriptional regulation." Methods **62**(1): 99-108.

Swain, P. S., M. B. Elowitz and E. D. Siggia (2002). "Intrinsic and extrinsic contributions to stochasticity in gene expression." Proc Natl Acad Sci U S A **99**(20): 12795-12800.

Tomancak, P., B. P. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, S. E. Celniker and G. M. Rubin (2007). "Global analysis of patterns of gene expression during Drosophila embryogenesis." Genome Biol **8**(7): R145.

Visel, A., M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin and L. A. Pennacchio (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers." Nature **457**(7231): 854-858.

Visel, A., E. M. Rubin and L. A. Pennacchio (2009). "Genomic views of distant-acting enhancers." Nature **461**(7261): 199-205.

von Ohlen, T. and C. Q. Doe (2000). "Convergence of dorsal, dpp, and egfr signaling pathways subdivides the drosophila neuroectoderm into three dorsal-ventral columns." Dev Biol **224**(2): 362-372.

Wartlick, O., P. Mumcu, F. Julicher and M. Gonzalez-Gaitan (2011). "Understanding morphogenetic growth control -- lessons from flies." Nat Rev Mol Cell Biol **12**(9): 594-604.

Weatherbee, S. D., G. Halder, J. Kim, A. Hudson and S. Carroll (1998). "Ultrabithorax regulates genes at several levels of the wing-patterning hierarchy to shape the development of the Drosophila haltere." Genes Dev **12**(10): 1474-1482.

Weingarten-Gabbay, S. and E. Segal (2014). "The grammar of transcriptional regulation." Hum Genet **133**(6): 701-711.

Weiss, J. B., T. Von Ohlen, D. M. Mellerick, G. Dressler, C. Q. Doe and M. P. Scott (1998). "Dorsoventral patterning in the Drosophila central nervous system: the intermediate neuroblasts defective homeobox gene specifies intermediate column identity." Genes Dev **12**(22): 3591-3602.

White, M. A., D. S. Parker, S. Barolo and B. A. Cohen (2012). "A model of spatially restricted transcription in opposing gradients of activators and repressors." Mol Syst Biol **8**: 614.

White, R. J. (2001). Gene transcription : mechanisms and control. London ; Malden, MA, Blackwell Science.

Wolpert, L. (1969). "Positional information and the spatial pattern of cellular differentiation." J Theor Biol **25**(1): 1-47.

Yanez-Cuna, J. O., E. Z. Kvon and A. Stark (2013). "Deciphering the transcriptional cis-regulatory code." Trends Genet **29**(1): 11-22.

Yao, L. C., S. Phin, J. Cho, C. Rushlow, K. Arora and R. Warrior (2008). "Multiple modular promoter elements drive graded brinker expression in response to the Dpp morphogen gradient." Development **135**(12): 2183-2192.

Yin, H., X. Xiao, X. Wen and T. Zhou (2013). "Stability of Regulatory Protein Gradients Induced by Morphogen Dpp in Drosophila Wing Disc." International Journal of Bifurcation and Chaos **23**(08): 1350138.

Zhang, C. C. and M. Bienz (1992). "Segmental determination in Drosophila conferred by hunchback (hb), a repressor of the homeotic gene Ultrabithorax (Ubx)." Proc Natl Acad Sci U S A **89**(16): 7511-7515.

Zhang, Y. T., M. S. Alber and S. A. Newman (2013). "Mathematical modeling of vertebrate limb development." Math Biosci **243**(1): 1-17.

Zhu, L. J., R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh, M. D. Basciotta, J. A. Brasefield, C. Zhu, Y. Asriyan, D. S. Lapointe, S. Sinha, S. A. Wolfe and M. H. Brodsky (2011). "FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system." Nucleic Acids Res **39**(Database issue): D111-117.

Zinzen, R. P., C. Girardot, J. Gagneur, M. Braun and E. E. Furlong (2009). "Combinatorial binding predicts spatio-temporal cis-regulatory activity." Nature **462**(7269): 65-70.

Zinzen, R. P. and D. Papatsenko (2007). "Enhancer responses to similarly distributed antagonistic gradients in development." PLoS Comput Biol **3**(5): e84.

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