DECIPHERING *CIS*-REGULATORY TRANSCRIPTIONAL GRAMMAR IN *DROSOPHILA MELANOGASTER* BY MATHEMATICAL MODELS

By

Ahmet Ay

A DISSERTATION

Submitted to
Michigan State University
in partially fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Mathematics

2009

ABSTRACT

DECIPHERING *CIS*-REGULATORY TRANSCRIPTIONAL GRAMMAR IN *DROSOPHILA MELANOGASTER* BY MATHEMATICAL MODELS

By

Ahmet Ay

Transcriptional regulatory information, represented by patterns of protein binding sites on DNA, comprises a key portion of genetic coding. Despite the abundance of genomic sequences now available, identifying and characterizing this information remains a major challenge. Minor changes in protein binding sites can have profound effects on gene expression, and such changes have been shown to underlie key aspects of disease and evolution. Thus, a key goal in contemporary genomics is to develop a global understanding of the transcriptional regulatory code, allowing prediction of gene output based on DNA sequence information. Recent studies have focused on endogenous transcriptional regulatory sequences; however distinct enhancers differ in many features, including transcription factor activity, spacing and cooperativity, making it difficult to learn the effects of individual features and generalize them to other *cis*-regulatory elements. We have pursued a unique "bottom up" approach to understand the mechanistic processing of regulatory elements by the transcriptional machinery, using a well defined and characterized set of repressors and activators in *Drosophila* blastoderm embryos. The study is concentrated on a set of proteins known as short-range repressors such as Giant, Krüppel, Knirps and Snail, which play central roles in development.

We have generated a large quantitative data set using fluorescent Confocal Laser Scanning Microscopy (CLSM) to determine the inputs (Giant, Krüppel and Knirps

protein levels) and outputs (*lacZ* mRNA levels) of the regulatory elements introduced into *Drosophila* by transgenesis. In this study (Chapter 2) we present a semi-automatic algorithm to process the image stacks from CLSM to correlate the protein levels of the short-range repressors with *lacZ* mRNA produced by reporter genes using images of *Drosophila* blastoderm embryos. We show that signals derived from CLSM are proportional to actual mRNA levels. Our analysis reveals that a suggested parabolic form of the background fluorescence in confocal images of early *Drosophila* embryos is evident most prominently in flattened specimens, with intact embryos exhibiting a more linear background. The data extraction described in this paper is primarily conceived for analysis of synthetic reporter genes, but the techniques are generalizable for quantitative analysis of other engineered or endogenous genes in embryos.

Using fractional occupancy based modeling on this data set (Chapter 3); we identified quantitative values for parameters affecting transcriptional regulation in vivo, and these parameters are used to build and test the model. We uncovered previously unknown features that allow correct predictions of regulation by short-range repressors on synthetic and endogenous elements. These features include a nonmonotonic distance function for quenching, which implicates possible phasing effects, a modest contribution for repressor-repressor cooperativity, and similarity in repression of disparate activators. This work provides essential quantitative elements of a transcriptional grammar that will allow extensive analysis of genomic information in *Drosophila melanogaster* and related organisms. Extension of these predictive models should facilitate the development of more sophisticated computational algorithms for the identification of *cis-regulatory* elements and the prediction of the quantitative output from novel regulatory elements.

Dedicated to my complaisant mother (Penbe Sahin),
my father (Bayram Sahin) and to my lovely wife (Ayten Ay)

ACKNOWLEDGEMENTS

So many people have touched my life throughout my time at Michigan State University. I would first and foremost like to thank my advisors, Dr. Chichia Chiu and Dr. David N. Arnosti, for their guidance, support, and encouragement over the last few years. Thanks to my committee members, Dr. Shin-Han Shiu, Dr. Zhengfang Zhou and Dr. Guowei Wei for their helpful comments about my research.

Where would I be without the fun, support, friendships, and frustrations that come with being an Arnosti group member? Thanks to all along the way who helped shape the scientist and the person that I have become. Thank you (alphabetically, of course) to Carlos, Jacob, Li, Liang, Lyle, Pankaj, Rupinder, Sandhya, Tess, Walid and Yang. I have been fortunate to have a wonderful group of friends at Michigan State, without whom, I am not sure I would have survived.

Not to be forgotten are the friends in my life that saw me through the process. Their support was invaluable. You helped me keep a healthy perspective about what life is really all about. Thanks so much. I would also like to extend a very special thanks to my wife Ayten for all of her love, support and patience during these years. I truly could not have done this without you.

Finally, and most importantly, I would like to say thank you to my parents. I would not be where I am today without your encouragement, your love, and your support. Thank you for always being there for me.

-Ahmet-

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Images in this thesis/dissertation are presented in color.

Chapter I

INTRODUCTION

Recent improvements in experimental techniques in biological systems have lent a new momentum to mathematical modeling in recent years. These changes have impacted studies of gene regulation the essential biological process that comprises the production of mRNA from the DNA and translation of mRNA into the protein. Gene regulation is a highly regulated dynamic process in which subtle changes, such as an increase or decrease of regulatory protein levels, can have profound effects. Such effects are involved in many human diseases such as cancer, as well as population differences and the evolution of morphological novelties. Despite an expanding knowledge based knowledge base on the biochemistry of gene regulation in the last fifty years, we still lack a quantitative understanding of this process. Mathematical models of the gene regulation offer an alternative approach for understanding the rules of this fundamental problem.

Most studies of gene regulation up to now have been experimental, constructing a "parts list" for transcription rather than generating the whole picture. These studies have provided an important but a qualitative picture of regulation instead of a physical and mechanical view. Now large scale sources of biological information are being generated providing a quantitative basis for systems biology studies. These include the complete genome sequences for many organisms, identification of many molecular factors involved in the regulatory processes inside the nucleus, expression levels for many genes measured at different time points, and in vivo occupancy of the DNA by DNA binding proteins such as transcription factors and chromatin modifiers. However, the quantitative understanding about how gene regulation works is far from comprehensive since the

created data usually gives only a snapshot of the system and obtaining this quantitative understanding with only experimentation is challenging. Mathematical modeling suggests an alternative path for this key problem. Today, with advances in molecular biology, genetic manipulation and the availability of complete genome sequences, development of new models that incorporate detailed dynamics of sets of biochemical interactions is possible.

Mathematical modeling has been used for understanding biological systems for decades (Turing, 1952). Although, these early models were breakthrough for their time and generally able to validate the experimental observations but could not provide any new insights on how the system works beyond which the experimenters had already proposed to be true. Recently the interest in this field exploded due to the new experimental and computational power. The new experimental techniques as mentioned above have provided a surge of biological data, which are difficult to understand without quantitative analysis and the improvements in the computational power, which is due to progresses in computational techniques and technology, enabled the calculations to be done which was not possible before. These improvements made the use of diverse mathematical modeling methods to different biological problems in particular gene regulation possible.

Systems biology provides a fresh perspective on gene regulation in biological studies and requires close collaborations between experimentalists and modelers. To carry out effective studies, the researchers who are involved in these collaborative studies in gene regulation need an understanding of both biological and mathematical fields. Here by focusing on the models that have been developed for eukaryotic systems in

particular yeast and *Drosophila*, we introduce the most common modeling approaches and their applications and focusing on goals, challenges and future directions. We also discuss how understanding of modeling can facilitate the understanding of eukaryotic transcription.

In modeling gene expression there are two general approaches. In the first approach, using largely statistical treatments, expression levels for hundreds of genes are analyzed. Here, gene expression levels under different experimental conditions or at different time points are used to find groups of genes that function together, new motifs for transcription factors or, to deduce regulatory relations between those genes. The ultimate goal of this approach is to construct the regulatory network that underlies the given data. Although this approach cannot explain the fine details of complex relations between transcription factors, RNA polymerase and other regulatory proteins, it can be very insightful on seeing the big picture, as it covers large fractions of genes in the organism. The second approach, which often involves analytical methods, describes the gene expression of a small number of genes, usually focusing on a detailed level of transcription using mathematical models. The models might include terms relating to binding of transcription factors and RNA polymerase to the DNA, cooperative and inhibitive interactions between transcription factors, mRNA and protein degradation, export to cytosol, and mRNA translation rate. For this approach we need an extensive level of knowledge and some hypotheses to check the system. In this review we will concentrate mainly on the second approach and discuss the use of major modeling techniques of this approach on gene regulation with examples from eukaryotic systems. We leave the discussion of the first approach to the other reviews in the field.

The models of the first approach such as graph based probabilistic models give a simple view of gene regulation; transcription factors bind to their target genes, and once bound, the factors can either stimulate or inhibit transcription. Those models could be used for analyzing the gene expression data coming from hundreds of genes, suggesting new regulatory relations that has not been discovered from experiments yet, finding feedback regulations in the system and comparing gene regulatory networks of different species, which might reveal conserved relations. However these models cannot be used to decipher the fine details of the system such as enhancer architecture and network dynamics. Neural, Boolean and Bayesian networks are examples of this approach and in this review as mentioned above we will not concentrate on them (Heckerman 1998; Friedman *et al*, 2000; de Jong 2002). In this approach the gene expression levels, regulatory factors like protein concentrations and experimental conditions might be used as input to the model.

A variety of mathematical models have been applied for quantitative understanding of gene regulation in eukaryotes, including thermodynamic models, Boolean models, differential equation models, and stochastic models. Those models are used to summarize the experimental data (Yuh *et al*, 2001), to infer new relations from complex experimental data, guiding the researcher to new hypotheses to test (Jaeger *et al*, 2004) and to find properties of the system that are hard to measure experimentally and can not be found without modeling (Fakhouri *et al*, 2009). These modeling approaches could be primarily categorized into four ways: discrete vs. continuous and stochastic vs. deterministic. In discrete models, time, state or space will assume a discrete set of values and in continuous models continuous values. On the other hand, deterministic models

describe the system from only the available experimental data, but stochastic models use probabilistic considerations to describe the system. All of these strategies have their strong and weak points, which we will discuss in detail in the following sections. For instance, a differential equation model, which is a continuous approach, serves as a good approximation for the underlying biological system, however it usually cannot be solved analytically, and numerical solutions are generally computationally expensive.

The choice of a model are usually depends on the system and problem under consideration. Several criteria should be applied in selecting a modeling approach; the model chosen should give new biological insights that could not be found with present experimental techniques or clarify some of the known connections; it should not only recapitulate what is already known. Although the aim of modeling is replicating the physical interactions of molecules, the complexity of the biology or lack of data for most processes obstructs us from doing that and cause most of the models in biology to be phenomenological.

Mathematical models of gene regulation depend heavily on the availability of good quality data, which until recently was largely available only in prokaryotic systems and yeast. Some aspects of gene regulation learned from these systems are applicable to more complex higher eukaryotic systems, but some features of gene regulation rules in higher eukaryotes are unique to these systems, such as the extensive use of alternative initiation sites and regulation at a distance. *Drosophila* has emerged as a model system for understanding the gene regulation in higher eukaryotes due to its elegant genetics, functional genomics (12 *Drosophila* genomes and genomes of other distant relatives are annotated), wide spectrum of methods for manipulations of the transcription, ease and

affordability of experiments as compared to other multi-cellular eukaryotes, and well established knowledge of the system due its long tradition. In addition, the results of the *Drosophila* studies are transferable to other eukaryotes such as humans: of 287 known human disease genes, 197 have homologs in *Drosophila.*

The *Drosophila* blastoderm embryo provides an ideal setting for the analysis of transcriptional enhancers; the cascade of maternally and zygotically supplied transcription factors has been extensively investigated at a molecular level, and many DNA regulatory elements have been identified and functionally dissected. In *Drosophila*, genes with complex expression patterns are controlled by multiple enhancers, whose modular function depends on the local action of repressor proteins (Small *et al*, 1993). As a result of the availability of high quality data sets the blastoderm embryo has been used for mathematical modeling of gene expression by several approaches, including reaction diffusion, Boolean, and fractional occupancy modeling (Jaeger *et al*, 2006; Sánchez & Thieffry, 2001; Segal *et al*, 2008). *Drosophila* segmentation networks' dependence on transcription for regulating its gene expression, the availability of enormous experimental knowledge on the system and available bioinformatics techniques for completing the unknown parts of the system makes this system unique for modeling transcription in higher eukaryotes.

In the following sections we will elucidate the applications of mentioned models particularly to *Drosophila melanogaster* and yeast. Our analysis suggests that although the mathematical models have been used to decipher gene regulation rules in higher eukaryotes, successfully.

THERMODYNAMIC MODELS

A general feature of most transcriptional elements in prokaryotes and many elements in eukaryotes is the activation or deactivation of a gene in response to binding of sequence-specific transcription factors (TFs). That is the occupancy of a regulatory element by TFs can provide a good proxy of the mRNA levels expressed by the associated gene. Although this picture ignores additional processes such as chromatin structure and modification, DNA methylation, and recruitment of general transcription machinery, the application of so called thermodynamic models permit the modeling of gene expression as a function of TF binding to promoter or distal enhancer elements. An underlying assumption for these models is that the level of gene expression is proportional to the equilibrium probability that the gene will be in an "active" state that is proportional to the number of activators, and inversely proportional to the number of repressors bound to the gene. The models seek to predict how different combinations of binding sites on a regulatory region function together to give diverse temporal and spatial expression outputs. These models are based on simple biophysical descriptions of DNA-protein interactions and statistical physics, and enumerate all possible 'states' of an enhancer based on potential transcription factor-DNA interactions. The probability of a gene firing is calculated as the fraction of the 'successful' states, i.e. those with preponderance of activators bound (Bintu *et al*, 2005a; Janssens *et al*, 2006; Zinzen *et al*, 2006; Segal *et al*, 2008).

An important step in modeling transcriptional regulation by thermodynamic models is connecting different states of enhancers to expression. All possible states of the enhancer are listed, and then a statistical weight for each state is assigned. For a simple case of one binding site, there will be just two states, bound and unbound. For an element

with four sites, there will be sixteen states etc. The statistical weight for a state is calculated by using the concentration of transcription factors and binding affinity of these factors to their binding sites on the DNA. Thus for abundant proteins binding to high affinity sites, the weight will be much greater than cases where the TF is scarce or the binding site is weak. After the weight for each state is calculated, the probability of each state can be calculated by dividing the statistical weight of the state by the sum of the statistical weight of all possible states. This calculation process can incorporate properties known to affect transcription. For example, cooperative interactions between transcription factors and inhibitory effects of repressors on activators can be explicitly added to the model by assigning higher or lower weights. These cooperative and inhibitory effects can furthermore be modeled to allow for distance effects. Competitive binding can also be incorporated to disallow simultaneous occupation of the same binding site by two different factors. After binding states and probabilities are defined, the next step in thermodynamic modeling is calculating gene expression output from each state. States with high activator occupancy are likely to induce high expression, while repressor occupancy might result in low expression. As discussed below, one can model gene expression output as proportional to the binding probability of the RNA polymerase or weighted sum of the transcription factors (Bintu *et al*, 2005a, 2005b; Segal *et al*, 2008; Fakhouri *et al*, 2009).

The theoretical underpinnings of thermodynamic modeling have been explored first and foremost in prokaryotic systems. Because the regulatory regions are generally small, binding to few TF, simple prokaryotic systems provide a tractable setting for quantitative studies and fractional occupancy modeling. The *lac* operon in *E. coli* and the

lysis/lysogeny switch of phage lambda are two examples that have been treated (Von Hippel *et al*, 1974; Ackers *et al*, 1982; Shea & Ackers 1985; Vilar & Leibler, 2003); reviewed in Buchler *et al* (2003) and Bintu *et al* (2005a). Additional promoters and configurations are considered in Bintu *et al* (2005a & 2005b). Zhou and Su generalized the results of Bintu *et al* (2005a) to derive a single formula calculating transcriptional probability for all simple regulatory configurations. The model is available as a Python module, 'tCal', which allows the user to easily build and configure transcription models of target genes.

The mechanisms of gene regulation found exclusively in higher eukaryotes suggest that TF binding to these regulatory regions may invoke distinct activities not captured by thermodynamic modeling of prokaryotic elements. However, a recent study by showed that, using thermodynamic modeling, many eukaryotic transcriptional responses can be executed by simple combinatorial logic. For example the Boolean AND gate can be generated by two TFs if the binding energies and factor concentrations are such that individual binding is weak, but joint binding is strong (Buchler *et al*, 2003). Larger groups of binding sites can encode a whole repertoire of more complicated logical functions. Their study showed that even in the absence of complex protein-protein interactions, by only changing the arrangement of binding sites, cooperativity of transcription factors and affinity of binding sites, complex interactions could be created (Buchler *et al*, 2003). One possible limitation to implementing some schemes may be the slow kinetics of assembling very large molecular complexes. Another prediction from these studies was that some complex regulation requires looping between distant sites and weak glue like interaction between proteins. The density of bacterial genome and lack of

chromatin boundaries may mitigate against extensive use of such long-range effects in prokaryotes, whereas higher eukaryotes have mechanisms to prevent intergenic cross talk.

Among all modeling approaches, thermodynamic models appear to provide the highest potential to predict the output of different combinations of transcription factor binding sites in eukaryotes. Recent studies involving yeast or *Drosophila* regulatory elements illustrate the possibilities and limitations of this approach in higher eukaryotes (Granek & Clark, 2005; Janssens *et al*, 2006; Zinzen *et al*, 2006; Segal *et al*, 2008 Gertz & Cohen, 2009, Fakhouri *et al,* 2009). In these models, parameters include cooperativity between proteins and the binding affinity of transcription factors to the DNA. Such models do not explicitly model other events such as chromatin modifications, RNA polymerase phosphorylation and promoter release however. The success of thermodynamic models suggest that those events which are downstream of the primary DNA and protein interactions might therefore play minor roles in the relationship between enhancer architecture and gene expression. Below, we discuss recent uses of thermodynamics models to study eukaryotic systems, and consider successes and limitations of these approaches.

A simplest method for transcription factor target detection is looking for single consensus binding sites in overrepresented binding sites to check whether they are regulating a group of genes. Such an approach has problems: motifs not the same with consensus sequence but close to it is not taken into account and often regulated genes have more than one binding site for a given factor. Granek & Clarke 2005 used thermodynamic modeling for detection of transcription factor targets in yeast genome by

using sequences, PWMs and concentration of transcription factors. They designed a transcription factor target detection algorithm (GOMER) which is unique in its ability to model competitive and cooperative interactions between transcription factors in comparison to machine learning methods. They applied their algorithm to yeast, but clearly it could be applied to other regulatory proteins. They used GOMER to identify Fkh2p and Mcm1p targets in controlling the expression of a set of cell-cycle regulated genes in yeast, and analyzed the role of cooperativity in this process. They also used the model to investigate the role of competition between the transcription factors, Ndt80p and Sum1p, in distinguishing between mitotic and meiotic programs of gene regulation. Another place they used their algorithm is predicting genome-wide binding of transcription factors, they guessed where the protein Rap1p is binding in the genome and then used chip-array to compare this model's prediction.

An alternative application of thermodynamic modeling in *Saccharomyces cerevisiae* involved intensive investigation of a *cis*-regulatory grammar applying to just the specific activator and repressor proteins, which are known to co-regulate genes in that organism. Cohen and colleagues constructed a set of over 2800 promoters containing three or four binding sites for these proteins. Quantitative output of each promoter was assayed by means of the fluorescent reporter, and the activities were fit by a thermodynamic model (Shea & Ackers, 1985). Their model could describe cooperativity between transcription factor binding sites, effects of weak binding sites, and the large amounts of variation in the gene expression due to different promoter architectures correctly. They used their model on the whole *Saccharomyces cerevisiae* genome to predict novel targets for the transcription factors they used. For example they found new

targets of Mig1 transcription targets, which was not identified bioinformatically previously due to their low binding affinity score.

At the other end of the spectrum, thermodynamic modeling has also been applied to a single, complex regulatory region to discover the detailed functioning of the element. Reinitz and colleagues modeled the activity of the promoter proximal 1.7 kb region the *Drosophila melanogaster* of even-skipped (*eve*) gene, which is expressed in seven stripes in the embryo. This region contains the modular enhancer directing blastoderm expression of stripe 2, as well as weak expression of stripe 7. After careful observation of the expression directed by this fragment, the authors incorporated the spatial and temporal expression levels of TFs regulating this gene into a thermodynamic model. Using only experimentally determined binding sites they were unable to recreate the expression patterns produced by reporter gene. However, when they included the likely binding sites predicted by bioinformatics techniques the model was able to fit the data. This study indicated that widely dispersed binding sites may operate together to generate enhancer-like outputs, suggesting that not all developmental elements exist as compact modules. Smaller segments of DNA may play necessary but not sufficient roles in driving gene expression. They extended their analysis by testing in silico the effects of mutation of specific binding sites or loss of specific transcription factors. The model was able to reproduce the altered patterns induced by these mutations. One limitation to this model is that it cannot be readily extended to other enhancer regions.

Segal and colleagues conducted a study that took advantage of high quality quantitative data available in the *Drosophila* blastoderm embryo extending the approach of Reinitz to 59 different enhancers. The model utilized spatial expression data for eight

transcription factors and expression of the target genes in mid-blastoderm embryos. Their model incorporated parameters for concentration scaling, homotypic but not heterotypic cooperative binding and expression contribution for each transcription factor. Unlike the *eve* promoter study, this model made no attempt to incorporate the distance effect of repressors, a critical feature of these proteins. Despite the simplifications of this model, reasonable predictions are obtained for many of the enhancers. The study predicted that weak binding sites make important contributions, as do homotypic interactions. This study highlighted the possible importance of weak binding sites to enhancer activity; these sites would provide a quantitatively significant level of activity and buffering against loss of single binding sites. This idea has yet to be validated, however. The contribution of homotypic cooperativity was also noted; this interaction provides sharper patterns at lower input concentrations. Their model generally predicts the expression patterns of the earlier expressed gap genes well, but is less successful with later expression patterns of pair-rule genes, possibly because of heterotypic cooperative interactions and distance-dependent quenching are not considered (Zinzen *et al*, 2006). They used ROC curves for evaluating the ability of their model to predict the expression patterns of modules that were not used as training data. Their analysis shows that their model is not over fitting the data; when they randomly generate PSSMs, swap PSSMs, permute the columns of PSSMs, they get worse results.

A distinct approach to thermodynamic modeling was taken by Papatsenko and colleagues, who focused on gene regulatory rules relevant to neurogenic gene expression in *Drosophila melanogaster*. The *rho*, *vnd* and *vn* are mainly regulated by two transcriptional activators Dorsal (Dl) and Twist (Twi), and one repressor Snail (Sna).

Differences in the regulatory regions for these genes lead to slight differences in expression patterns in dorsal and ventral regions. What was unique about this study is that rather than basing their thermodynamic model on the actual DNA sequences, the authors generated conceptual regulatory elements containing key core blocks of Dorsal-Twist-Snail sites. Their model was able to reproduce the relative gene expressions for *rho*, *vnd* and *vn* and suggested that the structural features such as cooperativity between transcription factors and Dorsal-Twist-Snail (DTS) module numbers could explain the expression differences. They concluded in their parameter comparisons that *rho* models require 5-10 fold higher Dl-Twi cooperativity than *vnd*, as well as higher Twi-Twi cooperativity, and *vnd* models require more DTS modules than *rho* as well as higher Sna-Sna cooperativity. Phylogenetic comparisons were employed to validate these conclusions. Comparing enhancer sequences from *Drosophila* species they noted that spacing between factor binding sites is generally conserved, and the number of DTS modules in *vnd* is always more than in *rho* so that features distinguishing *vnd* from *rho* have been maintained. They extended their study on *vnd*, mutating Twist binding sites and showing that their model can reproduce the experimental data. The actual function of modeled DNA sequences is not directly tested; however, most of the results of this paper were confirmed by earlier qualitative studies (Szymanski *et al*, 1995; Ip *et al,* 1992).

A combination approach was employed in a recent study that focused on modeling synthetic elements in silico using actual in vivo quantitative data obtained for each of the constructs (Fakhouri *et al,* 2009). Here 27 synthetic enhancers were devised to test features affecting in early *Drosophila* embryos. Levels of reporter gene activity were measured by confocal laser scanning imaging of over 900 embryos, and quantitative

differences resulting from minor changes in enhancer structure were noted. To obtain a fine scale understanding of the systems, only specific features affecting repressors were systematically explored. As explained above, earlier studies incorporating widely disparate enhancer sequences may not produce sufficient data to identify important features of enhancer organization, such understanding is critical to describe how extensively reorganized enhancers maintain similar function in some instances, or show quantitatively distinct outputs in other cases (Ludwig *et al*, 2005; Crocker *et al*, 2008). Arnosti and colleagues limited the number of features that differed between the relatively small number of reporter genes, enabling a model with a tractable number of parameters and robust estimation of these parameters. This study identified nonlinear quenching effects of short range repressors, similar quenching of different activators, and modest levels of cooperativity between short range repressors. Significantly, this type of modeling provided insights that were not apparent from the analysis of individual embryos. The study was extended to an endogenous enhancer, *rho* NEE, showing that certain features derived from synthetic enhancers are directly applicable to real enhancers and, highlighting several features of the architecture of this enhancer.

In considering these applications of thermodynamic models there are limitations and challenges that face the modeler. Eukaryotic transcription could be divided into three layers; binding of transcription factors, recruitment of cofactors and reduction of energy barrier of transcription by cofactors. This process is represented by three steps in the model of Janssens *et al* (2006): fractional occupancy of transcription factors and correction of activator occupancies by short-range repressor quenching, recruitment of cofactors (they use the adapter term in their study) and calculation of transcription rate by

Arrhenius mechanism. In the first layer of their model transcription factors bind to the DNA independently i.e. no cooperative binding, and occupancy of activators is corrected by reduction due to short-range repressor quenching effects. In this step quality of repressors is taken into account as free parameters, repressor quenching efficiency is assumed to be decreasing monotonically; complete repression up to 50 bp, linear decrease in repression 50 bp to 150 bp and complete loss of repression after 150 bp, and repressors is assumed to be cooperating on quenching activators which is represented by multiplication of their effects. The second layer of their model describes the cofactor recruitment by transcription factors, which is only a crude simplification of the process, where each activator has a constant potential to recruit cofactors. They also incorporated direct repression (reduction of RNA polymerase binding due to repressors) in an earlier version of their model in the second layer by decrease in cofactor levels; which is not incorporated by any of the models we mention here. Another simplification taken by them in this step is the assumption of activators recruit same cofactor, but biologically each activator might be recruiting different cofactors. The third layer in their model describes activation of transcription by Arrhenius mechanism in which cofactors lowers the activation energy barrier. In this step due to Arrhenius mechanism formulation, model gives cooperative effects between activators and nonlinear response to activation signal, and an exponential increase as more cofactors are recruited. To solve the exponential increase in transcription problem, a maximum threshold level is set, which does give a limit on transcription but not in natural way like logistic functions. Zinzen *et al* (2006) models only the first layer of transcription similar to Janssens *et al* (2006), and assumes that transcription level is assumed to be linearly correlated to the level of active states; in

his case the binding of activators but with at least one Dorsal activator and one Twist activator bound but not Snail repressor. In contrast to Janssens *et al* (2006) they have cooperative binding for transcription factors. Models of Granek & Clarke (2005) and Gertz & Cohen (2009) are similar to Zinzen *et al* (2006)'s implementation with some additional features such as weight functions for cooperativity. Segal *et al* (2008) modeled the transcription in three layers in a similar fashion to Janssens *et al* (2006); occupancy of transcription factors, summation of expression contributions of transcription factors and calculation of transcription by a sigmoidal function. The first layer of the model is similar to Janssens *et al* (2006) with the incorporation of homotypic cooperativity to occupancy calculations. In the second layer he parameterized expression contributions of different transcription factors, summed the expression contributions to get activation potential for each state. In the third layer he got the total transcription level by summing up the product of probability of seeing states and their expression contributions, which is calculated by using sigmoidal function. Model used in Fakhouri *et al* (2009) has the same set up with Segal *et al* (2008), but incorporates short-range repression and heterotypic cooperativity.

The gene regulation mechanism for eukaryotes is still not completely known and perhaps not all regulation mechanisms are found yet. Although we can incorporate all of the features known about gene regulation to the models quantitatively, we generally don't have enough data to test them. Here we will mention a few of those properties, which should be incorporated to the modeling as the data emerge. All the modeling attempts we mention here are only crude approximations to the truth and one might wonder how

realistic a model is, if it cannot mimic the complete picture of regulation. There are several key points to be noted in that respect.

First, although thermodynamic models could be incorporated to differential equation models, this incorporation is generally skipped and equilibrium assumption is taken i.e. the probability of molecules binding to the DNA is its equilibrium probability. Although the success of these models suggests that this is a reasonable simplification, it is still unclear whether these systems come to equilibrium and if they come how they manage that.

Second, currently most of the thermodynamic model applications could not take into account many structural features of the DNA in a realistic manner, especially for eukaryotic systems. For example effects of nucleosomes on transcription factor binding, distances between transcription factors, orientation of binding sites, closeness to transcription start site, chromatin modifications, DNA looing and so on. Although there are not much data available to add three dimensional picture of the DNA into the model, it is also not clear how most of these properties could be added. The simple cases of this such as looping could be incorporated into the model possibly by modifying statistical weights of states. Perhaps, the experiments in purified *in vitro* systems might be used for directly measuring the quantitative effects of these components such as chromatin modifiers and these effects could be incorporated to thermodynamic models also.

Third, although thermodynamic models is advantageous due to their rational and careful quantization of gene expression under different combinatorial states of transcriptional factors, they sometimes give up the dynamics of the system, and only concentrate on a snapshot of the system. These models could still reveal some useful

information about the system; however, we should note that steady state approach isn't as useful as the dynamic approach for modeling gene regulation due to the essence of biology which changes in concentrations matter a lot.

Fourth, in spite of all modeling studies, understanding the functional consequences of the changes in transcription are still a challenge. Although DNA based models might be one of the key contributors for answering this question, a network view of the system is needed for a clear explanation. The modeling attempts which help us to understand the fundamental rules between TFs might be used for constructing those network level models.

Fifth, eukaryotic enhancers in contrast to prokaryotes; might involve hundreds of binding sites and for those binding sites we don't know whether a certain binding site is functional at a certain time or not. Generally it is assumed if the protein is present, binding will occur and it would be functional, but clearly this is not the case for many binding sites. The lack of knowledge in where the proteins are binding and if they are binding are they functional is another key problem of modeling gene regulation, especially in eukaryotes.

Sixth, proteins competitively bind to DNA if their binding sites are overlapping and for some regulatory networks, competition is a key regulatory mechanism. Segal *et al* (2008) argued that competition for binding is not a major factor in gap gene network where, Zinzen *et al* (2008) argued the opposite in *vNE* genes. Although people use the assumption of overlapping binding site for competitive binding, due to steric hindrance constraints transcription factors with close binding sites might be also in competition for binding. This entity is taken into account by Granek & Clarke (2005); their competition

term incorporates all potential competitors binding at any window that affects binding of another protein. They program also allows the use of user defined weight functions for competitive binding.

Seventh, in thermodynamic modeling cooperativity is generally assumed at the level of binding, however could be added for later steps of regulation also. It is generally assumed that transcription factors bind to the DNA cooperatively, but neither proteins which cooperate nor their mechanism are known. Lack og knowledge in the correct cooperativity parameter values forces researchers to either estimate this parameter from experimental data or to incorporate it to the model in a very simple way (Granek & Clarke 2005; Zinzen *et al*, 2006; Segal *et al*, 2008; Fakhouri *et al*, 2009). Simplifying assumptions taken include; cooperativity only occurs between neighboring and same type of proteins (homotypic cooperativity), and it decreases by a distance dependent function, such as normalized Gaussian function with mean 0 and standard deviation 50 (Zinzen *et al*, 2006; Segal *et al*, 2008; Fakhouri *et al*, 2009). These assumptions definitely cannot explain the complex cooperativity schemes. For example importance of heterotypic cooperativity in gene regulation is shown previously by many experimental studies, so although ignoring heterotypic cooperativity simplifies the calculations it is not biologically reasonable (Zinzen *et al*, 2006; Segal *et al*, 2008). For example Granek & Clarke (2005) showed that heterotypic cooperativity Fkh2p-Mcm1p dramatically increased their models ability to explain the regulation of forkhead-regulated genes in contrast to homotypic cooperativity Fkh2p-Fkh2p which had only little effect. Granek & Clarke (2005) also incorporated user defined cooperativity weight functions to their

model which makes it easier to explore more elaborate models of cooperativity; however the choice of weight function for eukaryotes is not an easy task.

Eighth, actual concentration levels of the proteins are not known for most of the transcription factors in eukaryotic systems. People generally use confocal microscopy data, which measures relative levels of proteins and mRNAs, but levels of different proteins and mRNAs cannot be compared due to differences in the antibodies used or light wavelength (Janssens *et al*, 2006; Zinzen *et al*, 2006; Segal *et al*, 2008, Fakhouri *et al*, 2009). Since the thermodynamic models heavily depend on concentration levels, this lack of knowledge causes a key problem for thermodynamic models. A way to approach this problem is assigning free parameters for scaling concentration levels, which could be estimated by parameter estimation techniques (Segal *et al*, 2008; Fakhouri *et al*, 2009). Although this approach is used frequently, its applicability is still questionable due to lack of their validation. The improvements in confocal microscope technology will definitely improve the quality of the data, and help us to solve this dilemma.

Ninth, although transcription shows sigmoidal pattern is an assumption used by many researchers frequently, it is not proved to be true generally. For a correct modeling of gene regulation, this behavior should be tested further by experiments which compare different activator and repressor levels to transcription levels.

Tenth, some of the transcription factors function in context dependent manner in their activity. For example Hb is known to be a repressor on MSE3 and believed to be an activator on MSE2 due to bicoid presence, but the context dependency of this protein is not taken into account in the recent studies; Hb is taken as an activator in Janssens *et al*

(2006) and a repressor in Segal *et al* (2008). This property should also be incorporated to the models as more experimental data becomes available.

Eleventh, transcriptional regulation could be changed either by changes on regulatory regions or transcription factors. Since the effect of little changes on the transcription factor might possibly affect many other transcription networks, evolution generally takes the second option and does little changes on binding sites such as their affinity, composition and arrangement when there is a need to change the transcription level of a certain gene. This makes detection of binding sites a key starting point for modeling transcription; however transcription factors can tolerate high sequence variability, which gives a high flexibility to gene regulation and makes the detection of binding sites a complicated task. The number of known binding sites experimentally is limited and bioinformatics techniques do not guarantee detecting binding sites precisely, but this makes the results of modeling studies prone to errors or over fitting. In Janssens *et al* (2006), use of 17 experimentally detected binding sites were not enough to fit the model to the gene expression data; they were only able to fit their model when 17 bioinformatically detected binding sites were added.

Twelve, although many modeling studies treat the relationship between the transcription factor and its target binding site as something that either does or does not exist, biophysically this does not make sense and a quantitative description which assigns a wide range of affinities to the transcription factor and its target binding sites on the DNA is needed. A theoretical outline for finding the relative binding affinities of binding sites is described by Berg & Von Hippel (1987). Their outline describes an approximate relationship between the Position Weight Matrix (PWM) score which is inferred from

base frequencies in known binding sites and the relative affinity of the binding site assuming additivity of the binding energy for each base pair. Although, if we mutate each bp in binding site one by one and check binding levels we can actually find free energy contributions for each bp, this is costly. Berg & von Hippel (1987) framework is thought be a nice approximation due to the assumed connection between the information content in the set of sequences and specificity of binding. This framework takes into account the assumption that natural selection has given rise to a certain level of sequence specificity for each TF and the sequences that give the same binding affinity are selected. From a biophysical point of view this formalism is not satisfactory, evolutionary arguments should not be invoked when the goal is to model the physical interactions between a transcription factor and the DNA sequence. In addition, although the strength of TF-DNA interaction varies with the local DNA sequence, it should not depend on the choice of a background model representing the global characteristics of the DNA. Purely biophysical approaches are needed to infer TF binding specificity; as an example, transcription factor binding specificity could gain a lot from structural information on transcription factor-DNA binding and this information might open up new possibilities for determining TF sequence specificity.

Thirteenth, generally the binding affinity is taken care of by use of PWMs or totally ignored and each binding site for a transcription factor is assumed to have the same binding affinity. Both of these ways have problems; on one hand there is not enough number of binding site data available for most of the transcription factors to construct a nice and representative PWM, on the other hand assigning the same affinity for each binding sites of a transcription factor is an oversimplification. In two recent

studies a third approach is taken and binding affinities are optimized for each binding site from the experimental data and they argued that this approach didn't improve their results significantly compared to the classical approaches mentioned above (Janssens *et al*, 2006; Zinzen *et al,* 2006). Zinzen *et al* (2006) also argued that a wide range of binding affinity values $7 < \log(K_a) < 11$ could not account for the measured differences in the *rho* and *vnd* expression patterns and they are not as important as cooperativity parameter for thermodynamic models. The insensitivity of gene expression to changes in binding affinity due to model is probably a weak point of thermodynamic models.

Despite all its shortcomings, thermodynamic models still stay as the most promising model for deciphering gene regulation on DNA level. Although we don't have sufficient data for some of the gene regulation features mentioned here, mathematicians should go forward and check the applicability of them on synthetic data and show new ways to experimentalists to create reasonable data sets for deciphering gene regulation. On the other hand, the incorporation of thermodynamic models to large gene regulatory network studies will possibly increase their performances and this opportunity should be also further investigated.

DIFFERENTIAL EQUATION MODELS

A dynamical system is a set of components that interact by explicit rules which dictate how the states of the components in the system change. One example of dynamic systems is regulatory networks in which the components are mRNAs and proteins, and the state of these components is their concentration levels. The rules of this evolving system of interacting molecules could be represented by differential equation models in terms of the rate equations for expression of any variable in terms of the other variables.

These models could incorporate parameters such as the degradation rates of mRNAs and proteins, or the firing rate of promoters. Importantly they could be used to reproduce observed system behavior or perform in silico tests of the system parameters at levels which are hard to get by experimentation, leading new hypotheses to check.

Differential equation models are based on the principal idea that there is a sufficient number of components that we can consider them as continuous quantities which are spatially homogeneously distributed. Although these assumptions are oversimplifications of the real problem, and usually do not hold true for biological systems-differential equation models still provide many useful insights about complex regulatory networks.

In general terms a differential equation model can be simplified to $\frac{dx_i}{dt} = f_i(x)$, $1 \le i \le n$ where $x = (x_1, x_2, ..., x_n)$ is a vector of concentrations of mRNAs, proteins or other related molecules and f_is are rate equations, which express the changes due to transcription, degradation, translation, etc. The selection of an f_i is a critical step in modeling and one should balance complexity with detail. Complex models are good to explain the system, but they will contain many parameters, which might be hard to measure experimentally or estimate computationally. Furthermore if necessary, f_is might be chosen to take into account the time-delays and uncertainty in the system.

Differential equation models can be divided into two main groups; Ordinary Differential Equation (ODE) models and Partial Differential Equation (PDE) models. ODE and PDE models depend on one or multiple variables respectively. For instance, for temporal changes only we use ODE models, for both temporal and spatial changes, we

use PDE models. In modeling gene regulation ODEs are the most common formalism. They include time dependent variables such as protein and mRNA concentrations, and constants for production and degradation rates for each variable in the system. Although PDEs are more appropriate for modeling gene regulation, the requirement in mathematical analysis and experimental data makes them less favorable.

ODEs are a well studied field of mathematics; the ODE theory is well established and many numerical methods and software tools for solving ODEs are freely available and easy to implement. Although, ODE models are generally hard to solve analytically, i.e. finding formulas that express the solutions as explicit functions, approximations of the solutions can be found by using numerical methods. Usually the spatial structure is not taken into account in ODE models, when it plays an important role in the system, however, an extra parameter is written into the model so that it could be taken into account. In this case, we consider the spatial structure as a collection of homogeneous compartments, between which information can be transferred. ODE models are insufficient for modeling gene regulation if continuous aspects of the geometry of the system are important or the homogeneity assumption is not acceptable to the system. PDEs should be used for these cases.

PDEs, like ODEs are well studied analytically and numerically. Unfortunately, its theory is more complicated and computations are more demanding. Finding analytical solutions are much harder, therefore numerical simulations are the main analysis tool for biological systems. Those models could take into account the spatial aspects of the cellular processes if necessary (Eldar *et al*, 2002), however, spatial heterogeneity of the systems is generally neglected. We will see the applications of a classical example of a

PDE (the reaction diffusion equation) below, which describes the production, diffusion and degradation of biological components.

Following standard methods of chemical reaction kinetics, one can obtain a set of differential equations for any regulatory network, which could be solved numerically to a high accuracy level using numerical methods such as Bulirsch-Stoer Method (Press *et al*, 1992). This standard modeling approach has been applied to many systems, ranging from a few isolated components to entire cells. This modeling approach has been previously applied in many prokaryotic regulatory networks such as *lac operon* of bacteria. The *lac operon* consists of a number of genes and a small regulatory DNA region, which controls the expression by binding to either repressor or RNA polymerase. The first mathematical model to study *lac* operon regulation was given by Goodwin in 1965 and Griffith 1968. However, the beauty of the problem attracted many other researchers in the past 50 years (Nicolis & Prigogine, 1977; Santillan & Mackey 2001; Yildirim & Mackey 2003; Mackey *et al*, 2004). For example, the model of Griffith takes into account activation of the genes, transcription of mRNA, degradation of lactose, synthesis of beta galactosidase and permease. Later, the Nicolis and Prigogine model added more details about the system, including the action of the repressor, inducer and enzyme synthesis. Yildirim and Mackey's model added delays in the system, such as transcription initiation and translation. One wonders whether all these attempts could be generalized to other DNA regions or not?

Although differential equations have been used abundantly by modelers for many biological systems such as population dynamics, embryo patterning, infection dynamics and gene regulation, the framework they provide is largely unknown to biologists. Here

we will briefly describe the use of differential equation models as a tool for describing and making predictions about temporal and spatial changes in eukaryotic regulatory networks, and discuss their strengths and weaknesses. Use of differential equation based models in eukaryotic regulatory networks is relatively new; however, the accumulated data on molecular biology such as the segmentation network in *Drosophila* makes the use of these models a preferable choice (von Dassow *et al*, 2000; Gregor *et al*, 2005; Jaeger *et al*, 2004). We will also present several recent applications of these models to facilitate testing hypotheses about systems level properties of complex eukaryotic systems.

Barkai and colleagues studied the network of proteins that pattern the dorsal region of the *Drosophila* embryo, which is initiated by the graded activation of the bone morphogenetic protein (BMP) pathway (Eldar *et al*, 2002). Their results indicate that the BMP activation gradient is robust to changes in gene dosage. However, the mechanism of robustness is different than the general design of robust networks, which involve feedback loops to buffer against perturbations in the system. Their modeling of the system suggested the transport of the Scw and Dpp, BMP class ligands, into the dorsal midline by Sog, a BMP inhibitor, as the key event in robustness. They validated this result experimentally for Dpp.

Three reaction diffusion equations that form their model are written for the system. Their model could account for the formation of the BMP-Sog complex, diffusion of Sog, BMP and BMP-Sog, and allowed for the cleavage of Sog by Tld, both when Sog is free and when Sog is associated with BMP. They carried out 66000 simulations to find the robust networks where they changed parameters of their model such as the rate constants and protein concentrations over four orders of magnitude and solved their

model numerically. They generated and solved numerically three perturbed networks representing heterozygous situations by reducing the gene dosages of sog, tld or the dpp (or scw) by a factor of two and compared their output with the initial, nonperturbed network. They found that only 198 out of 66000 networks are robust to twofold changes in three genes, which showed a unique sharp concentration gradient that peaked in the dorsal midline. Analysis of the parameters showed that the robust networks could have a wide range of possibilities for most parameters. However, there are two restrictions on the design of the network; cleavage of Sog by Tld is facilitated by the formation of the complex Sog–BMP and the BMP–Sog is broadly diffusible, while free BMP is not.

Maternal morphogen bicoid (Bcd) of *Drosophila* diffuses along the anterior-posterior axis of the embryo and forms a gradient in the early blastoderm stage of the *Drosophila* embryo. This gradient helps the formation of anterior-posterior patterning, which assigns different fates to nuclei depending on the level of Bcd they are exposed to. Gregor and colleagues wanted to understand how it is possible to get similar anterior-posterior patterning within different *Drosophila* species with different sizes of embryos (Gregor *et al*, 2005). In order to see how diffusion works for Bcd they injected an inert, fluorescently tagged molecule that mimicked the Bcd at the anterior pole of the embryo and measured the concentrations at different spatial points of the embryo over time. They modeled this process by reaction diffusion model, which describes the change in protein concentration pursuant to diffusion over space and time as well as decay due to protein lifetime. In order to solve their model numerically over time, they discretized the embryo into a three-dimensional grid. They fit the experimental data with their model for different species and found species-specific diffusion constants. In species with different

sized embryos the relative change in diffusion constants was minimal and also the difference in developmental time scales of the species was small. These results led them to hypothesize that the reason for the identical patterns over different *Drosophila* species is due to species specific lifetime of Bcd. However, this result has yet to be experimentally verified.

Shape and stability of Bcd morphogen are always assumed to be the result of localized production, diffusion and degradation. Recently, the diffusion rate of Bcd protein is reported but the rate of Bcd production and degradation remains uncertain (Gregor *et al*, 2005). On the other hand, it has been shown by recent live-imaging experiments that Bcd undergoes rapid nucleocytoplasmic shuttling and equilibrates between the cytoplasmic and nuclear compartments (Gregor *et al*, 2007). However, the equilibrium level changes in time due to the number of nuclei present to trap the Bcd. These recent observations led Shvartsman and colleagues to suggest a new differential equation model to explain the exponential shape of Bcd without degradation of Bcd protein; their model incorporates constant localized production at the anterior pole of the embryo, diffusion and nucleocytoplasmic shuttling in the presence of the growing number of nuclei instead of localized production, diffusion and degradation (features of earlier reaction diffusion models) (Coppey *et al*, 2007). Their model predicts that nuclei do not contribute significantly to the shape of the Bcd gradient; the Bcd gradient establishes before the nuclei migrate to the periphery of the embryo (blastoderm stage) and remains stable during subsequent nuclear divisions, i.e. nuclei can be viewed as essentially inert sensors of the pre-established concentration from earlier time points. The existence of the stable dynamics of the profiles of nuclear Bcd is a robust feature of the

model; the parameters of the model do not have to be fine tuned. Another prediction of the model is that local defects in nuclear density should generate only local defects in the profile of nuclear Bcd. Although these predictions have not been tested yet, they could be tested by careful measurement of Bcd stability and analyzing mutants with late defects in nuclear migration. Despite these nice predictions, their model cannot account for scaling of the gradient with the size of the embryo similar to earlier reaction-diffusion models (Gregory *et al*, 2005, 2007).

Anterior-posterior patterning in the early *Drosophila* embryo is one of the well studied biological systems experimentally by many enhancer bashing and gene mutation experiments. However, these studies are usually insufficient to get a complete picture of this patterning process. To get the complete picture experimentally, we need to create an in vitro system which reconstitutes the underlying gene network from well defined ingredients. This reconstruction is almost impossible for right now, and for this reason alternative approaches are needed. Therefore in recent years researchers have been trying a new approach using mathematical modeling to recreate this network in silico and test the hypothesis.

In several recent studies Reinitz and colleagues analyzed the *Drosophila* gap gene regulatory circuit (one hour prior to cellularization in *Drosophila*) by using reaction-diffusion models, which was used by the same group in an earlier study to analyze the formation of stripes f expression of the pair-rule gene *eve* (Jaeger *et al*, 2004a; Jaeger *et al*, 2004b; Reinitz & Sharp 1995). In their study they used spatial and temporal data of wild type protein levels of the network, which constituted of the gap genes hunchback (hb), Kruppel (Kr), knirps (kni) and giant (gt), maternal factors bicoid (bcd) and caudal

(cad) and the zygotic gene tailless (tll). Their model provides a way to use this data to infer how concentrations of products of a given gene change with time and how these changes are influenced by the activating or repressing effects of the products of other genes. Their model is based on three main ideas; protein concentrations are taken as state variables of the network, chemical reaction kinetics are given by coarse grained rate equations for protein concentrations, and least squares fit is used to estimate the parameters of the network. The optimized parameters suggested to them regulatory relations in the network, which they compared to the literature and checked if their model could successfully mimic the gap gene circuit and known mutations in the circuit. Their model agreed with the earlier mutant and reporter studies, recommended that some of the previously reported regulatory interactions are not necessary for getting the model to fit and suggested some new regulatory interactions such as activation of Kr by Cad.

They show that their model could reproduce gap gene expression at high accuracy and found the mechanisms previously inferred from qualitative studies of mutant gene expression data (Jaeger *et al*, 2004a, 2004b). Their analysis argues that threshold dependent interpretation of maternal morphogen concentrations is not sufficient to explain shifts in gap domain boundary positions. They argue that maternal factors start activation of gap genes but their fine detail will be determined by gap genes i.e. positioning of gap gene boundaries and maintenance of gap gene expression depend mostly on gap-gap talks rather than maternal activators (Jaeger *et al*, 2004a, 2004b). Their model also suggested diffusion as a non critical mechanism for observing gap gene shifts; if they don't allow shifts in the gap gene networks, they still see the shifts in gap gene expressions (Jaeger *et al*, 2004a).

The model predicts that synthesis is confined to the anterior region of each expression domain, which implies that there is an asymmetric distribution of gap gene transcript in protein domains; transcript domains of Kr, kni and gt are shifted anteriorly with respect to their corresponding protein domains. They argue that these shifts are due to protein synthesis domination anteriorly and protein decay domination posteriorly. Shifts of anterior gap domains could be considered secondary effects of shifts of posterior gap domains, i.e. shifts of anterior gap domain boundaries either follow the posterior boundary shifts of more anterior gap genes or are due to sharpening of posterior boundaries of anterior gt and hb (Jaeger *et al*, 2004a).

They predict activation of Kr by Cad and several other regulatory relations (clarify evidence on the effects of Hb on *Kr*, Kr on *kni*, and Gt on *kni*) (Jaeger *et al*, 2004b). Unfortunately, most of the results from this study have not been validated yet. It has also been noted that the model fails on null mutant experiments. They claim that this failure is due to indeterminacy in quantitation of protein signal, or due to false early gap gene regulation, which predicts high levels of gap genes that does not actually exist.

The periodic spatial pattern of segment polarity gene expression of *Drosophila melanogaster* along the anterior-posterior axis of the embryo is maintained throughout development, providing positional information for subsequent developmental events. Segment polarity genes process and maintain their expression state through cross regulatory interactions, which involves cell to cell interactions. Recently, von Dassow and colleagues analyzed the segment polarity network by ODE models and checked whether the known interactions in this network are enough to yield robustness (von Dassow *et al,* 2000).

They developed a logical network interpretation from experimental results for the reactions between the segment polarity genes and their products, including Engrailed (EN), Wingless (WG), Hedgehog (HH), Patched (PTC) and Cubitus interruptus (CID). They used this logical network for designing an ODE model that encodes this logic in a set of 13 nonlinear ordinary differential equations, which incorporates synthesis, decay, heterodimerization, cleavage and cell-to-cell traffic. The model encoded 50 parameters including binding rates, cooperativity coefficients and half lives of proteins and mRNAs where the real values are usually unknown. In their study, they tried to find a set of parameter values which the model exhibits the desired behavior of a segment polarity network, given realistic initial conditions. Their model could not result in the segmentation patterns observed with the known interactions among the segment polarity genes and their products in the observed behavior of the embryo during and after the segment polarity stage. Even after attempting a wide range of constrained parameters for their model, they could not determine one parameter set that would maintain the initial pattern of these genes' expression stably over time.

They argued that the discrepancy seen is possibly due to lack of information about the topology of the network rather than the parameters chosen. They redesigned the protein-protein interaction network with some additional interactions, which was suggested by their best approximate results; *wg* and *en* expresses alternately in every other cell along lines of cells parallel to the anterior–posterior axis. They added two new interactions, which are biologically reasonable to get the expected pattern; *wg* auto-activation, and inhibition of *en* by Ci amino-terminal repressor fragment. The difference between this new model and old model is the emergence of two positive feedback loops,

one for *en* and one for *wg*. With these links installed there are many parameter sets that enable the model to reproduce the target behavior, i.e. the segment polarity network is robust to parameter variation. Their model also suggests that the segment polarity network requires few absolute demands on initial conditions, and it seems likely that the evolutionary process could replace those inputs relatively easily which aids in the use of the same module in different organisms. They also showed that robustness property is not an artifact of the network topology; they analyzed models that include additional links and components and observed that as long as they keep the core topology the same conclusions hold. Although the results of this model haven't been experimentally validated yet, the approach itself was extremely valuable in guiding new experiments.

There are a few points to be noted in differential equation based models. First, three-dimensional structure of biological systems should be taken into account for realistic modeling. Although reducing the dimension of the system might simplify calculations and be a reasonable first approximation to the problem, models should be extended as new data emerges. For example, Jaeger *et al* (2004a, 2004b) and Reinitz & Sharp (1995) concentrated on a strip of cells from the middle of the *Drosophila* embryo with the assumption that anterior-posterior and dorsal-ventral patterning networks are independent in *Drosophila melanogaster*, but it is accepted that this assumption is not entirely true. Recently, Fowlkes *et al* (2006) and Luengo-Hendriks *et al* (2006) noted the importance of nuclear movements on gap gene regulation, which could also be taken into account in a three dimensional treatment of the problem. They are producing three-dimensional data on *Drosophila* segmentation genes, which the modeling approach taken by Reinitz and colleagues could be applied to (Fowlkes *et al*, 2006; Luengo-Hendriks *et*

al, 2006). Similarly, Coppey *et al* (2007) assumed in their model that nuclei are uniformly distributed throughout the embryo which contradicts also to studies of Fowlkes *et al* (2006) and Luengo-Hendriks *et al* (2006).

Second, although differential equation models are more suitable than other modeling approaches like Boolean models for the dynamic nature of biological systems, the quality and quantity of data needed to construct a differential equation model makes them difficult to apply. If there is insufficient data or the data of poor quality, use of these models might result in inaccurate predictions. However, the inaccuracy could be resolved by increasing quality and quantity of the data, and the differential equation models could be solved numerically to any desired precision. For example, although Reinitz *et al* (2004a, 2004b) used the same modeling and optimization techniques with Reinitz *et al* (1995), they had a lower degree of variation in the distribution of parameters and the error levels in the gap gene expression patterns were reduced to less than 5% by the increase in the data quality.

Third, although we assume that cells and cell compartments are homogeneous, in reality cell compartments are highly heterogeneous and compartmentalized structures. This leads to a situation where the discrete nature of the molecular components cannot be ignored, resulting in the stochastic behavior of biological systems. Although, differential equation models are appropriate for simulating systems which satisfy continuum approximation, they fail to describe stochastic interactions. Approaches that model stochastic effects between individual molecules offer better descriptions of system behavior in such cases, such as the Gillespie algorithm (Gillespie *et al*, 1977).

Fourth, differential equation models usually depend on lots of parameters that quantify the interactions between molecules and finding these parameters experimentally is not an easy task. Consequently, these models often become underdetermined and many parameter values could work equally well. Therefore, producing quantitative predictions for large biological systems by differential equations is not straightforward and this could only be done for simple systems. On the other hand, the numerical techniques for solving differential equations should be chosen carefully and the stability of the solutions should also be checked. In a numerically stable algorithm, errors in the input lessen in significance as the algorithm executes, having little effect on the final output, and in a numerically unstable algorithm, errors in the input cause a considerably larger error in the final output. Finally, due to high computational needs these models do not scale well to complex regulatory networks with hundreds of interacting molecules. However, since they protect the overall picture of the biological systems, as opposed to other modeling approaches such as Boolean models, the disadvantages in computation are compensated for by the accuracy of the results. This problem might be potentially solved by the improvements in computational techniques and technology.

Fifth, discrete time scale could be added to the biological systems, if necessary, by using difference equations, which is the discrete form of differential equations where the value of one variable at a certain time depends on the value of that variable at a former time. Difference equations are more realistic than differential equations due to the discrete nature of biology. Although the use of discrete time in modeling is correct for ecology models in which new organisms are born in synchrony, it is less appropriate when no natural time step exists, such as in transcription where transcription time varies

across genes. On the other hand, the time delays in gene regulation due to involvement of processes such as import and export to the nucleus or slowness of the processes such as translation could be incorporated into the model using delay differential equations.

Sixth, there is very limited information about the actual in vivo processes, and for this reason lots of assumptions are made about nonlinear biological processes such as spatial organization of the system, interactions between molecules, and reaction rates. For example, it is generally assumed that interactions among molecular species follow mass-action kinetics, but mass-action kinetics may not be suitable for some reactions such as conformational changes in large-scale macromolecular aggregates. On the other hand, in these models the DNA level regulation of transcription is taken as a black box, and usually approximated by a function of sigmoidal type, which may not be realistic since every protein does not have to obey the same type of sigmoidal behavior; some proteins may be extremely sensitive to increases in activation signal where others may not.

Seventh, it is believed that the regulatory networks are composed of modules, which are assumed to be semi-independent in their behavior. For example, in a typical *Drosophila* embryo it is assumed that the genetic regulatory networks are effectively isolated from other developmental processes such as cell-cell interactions, morphogenetic movements, and protein phosphorylation. Although differential equation models do a pretty good job at describing the behavior of the modules, due to possible missing pieces in the modules their use is limited; for example, changes in the modules such as addition or subtraction of new proteins might have profound effects, which may not be pointed out by differential equation models due to over fitting to the earlier description of the module (von Dassow, 2000).

Eighth, differential equation models usually ignore post-translational regulation and take DNA level gene regulation as a black box. For this reason, these models can't help us to understand the enhancer structure and organization, such as cooperative binding, distances between transcription factors, orientation of transcription factors; however, these models are useful for finding potential targets for DNA level studies. Because of skipping DNA level regulation, the rules learned from one circuit is typically not applicable to other circuits, since on the DNA level protein binding sites might be reorganized and this reorganization might change the rules learned on protein level. Although this is a weakness inherent in the studies that do not take into account the DNA level information, it does not mean that this approach is useless-it might really help us to see the bigger picture. However, if the rules on the DNA level are found they could be connected to the protein level analysis and possibly used by other circuits.

Ninth, the fitness of differential equation models are measured by cost functions, which compare the model's predictions to experimental data and visual inspection. There are two approaches to this problem; finding parameters of the system by an optimization method which minimizes the cost function or taking a parameter range for parameters and trying all possible values. Usually for parameter estimation global optimization techniques are used, which are known to give multiple parameter sets that can satisfy the system equally well due to their stochastic nature. For this reason interpretation of the parameters is not easy and clearly the most important part of modeling. For example, Reinitz and colleagues used Parallel Lam Simulated Annealing for estimating parameters, which resulted in many gene circuits which they checked for defects and used the circuits which do not have any observable defect for further analysis (Jaeger *et al*, 2004a, 2004b).

Alternatively, Eldar *et al* (2002) choose 66000 different parameter combinations where each parameter ranges over four orders, and they checked for the ones which give robust pMad expression.

BOOLEAN MODELS

Biological processes often show switch-like behavior such as competence in bacteria, apoptosis decision in cells and transcription of a gene. For this reason, Boolean models which represent the regulatory relations as logic gates are one of the most studied discrete approaches that could capture and describe this behavior. In this approach, mRNAs and proteins in the network are assumed to be binary valued logical variables, i.e. their states can be either on or off which is usually decided according to a threshold value. The associations between the variables are described by Boolean or logical functions, which provides a statement performing on the inputs, mRNAs or proteins that have regulatory signal to the target, using the logic gates such as "and", "or" and "not", and the output is on (1) or off (0). For instance a gene which is regulated by two transcription factors; AND function implies that the gene is transcribed only if both transcription factors are binding, OR function implies that the gene is transcribed even if one of the transcription factors are binding, and NOT function implies that the gene is not transcribed if both of the transcription factors are bound.

Boolean networks can be used for simulating dynamic behavior in biological systems by applying the Boolean functions in discrete time steps, usually the time interval that is larger or equal to the duration of all biological processes in the system. The updates in the network, which could be synchronous or asynchronous, cause the biological system to evolve from state to state, where a state of the biological system is a

binary vector demonstration of the states of each variable (mRNA, protein, etc.) in the system. In this approach each variable has two states-on (1) or off (0), and the dynamics show how the variables change each others' states over time. Due to the deterministic nature of Boolean models, the initial state of the system completely determines the end state of the system. If no difference occurs between transitions of states, then the system is in a point attractor state (analogous to steady state in differential equations), and if states of the system repeat periodically, then the system is in a cycling attractor state (analogous to limit cycles in differential equations).

Boolean modeling could be used for any biological system, where interactions between its elements are well described, to combine qualitative experimental observations in a logical structure. Due to their simple nature, they circumvent the need to know quantitative details about the reactions in the systems, which is not available for many biological systems. This simplification creates an advantage for Boolean models over other modeling approaches (such as differential equation models which usually include many unknown parameters), they become mathematically more tractable which makes them easy to analyze analytically and implement computationally. For example analysis of the steady states or limit cycles of the Boolean model is much easier than differential equation models. However, despite their simplicity, they could still provide qualitative insights into the fundamental nature of the underlying system, such as the role of feedback loops in the network's behavior. Also, since they are easy to analyze and implement, they could be extended to large scale biological systems with thousands of players.

Boolean network models could be used as an exploratory tool for systems where the network structure is not clear. Many variants of the same network could be created, and the simulations of the model could be compared to previous experimental findings and the instincts of the modeler. As a result, Boolean models could be a way for the modeler to understand the dynamics of the system and start modeling even without knowing fine details of the system.

Biological networks are usually robust despite intrinsic and extrinsic noise, for example protein and mRNA levels are noisy with varying activity, and dynamic. However, by adding stochasticity to Boolean models they could be extended and used for analyzing the conditions under which the biological network is robust. On the other hand, use of coarse-grained Boolean models to capture dynamicity in the networks seems unrealistic. Recently several studies used these models based on very simple Boolean functions to fit sequences of gene expression patterns (Sanchez *et al*, 2001; Albert & Othmer 2003). We will analyze these studies below for a further understanding of these models.

Recently, Albert & Othmer (2003) have constructed and analyzed a Boolean network model for segment polarity genes of *Drosophila melanogaster*. These genes show stable expression patterns which supply the necessary information for the following developmental processes. The expression of the segment polarity genes are refined and sustained throughout a number of developmental stages by regulatory interactions between the elements of the network, which constitute not only cellular interactions but also intercellular interactions. As mentioned in the previous section, analysis of this network by nonlinear differential equation models suggested that the steady-states of this

network are not affected by choices in the kinetic parameters and determined mainly by the type of regulatory interactions within elements of the network and topology of the network. Albert & Othmer (2003) constructed a Boolean network model which recapitulates the main conclusions of von Dassow *et al* (2000) and accurately predicts the dynamics of this network. Their model is based on binary ON (1) and OFF (0) representations of mRNA and protein levels in the network and the interactions between elements of the network are defined by logical functions. The Boolean network is updated every unit time step to create the next state of the network. They also changed their one step model to a two-step model, where they assumed that proteins degrade in two steps, but mRNAs degrade in one step. However this alteration in the network did not change the main conclusions of the model.

The model's performance is measured by its prediction of spatial and temporal gene expression levels of the network, which are present (1) or absent (0), rather than absolute continuous levels of mRNAs and proteins. In their model the following assumptions were made: the effect of transcriptional activators and inhibitors is never additive, but rather, inhibitors are dominant, transcription and translation are ON/OFF functions of the state, if transcription/translation is ON, mRNAs/proteins are synthesized in one time step, mRNAs decay in one time step if not transcribed, transcription factors and proteins decay in one time step if their mRNA is not present. In their modeling they take into account only 12 cells i.e. 3 parasegment primordia and impose periodic boundary conditions. 4 cell per parasegment primordium are used since when expression of segment polarity genes begins, a given gene is expressed every four cells. They did not add the nodes such as FZ and smo to the modeling since these are not regulated by other

nodes in the network. With all the elements they have, the total number of nodes becomes 180. They add more nodes which are composites of two or more, to make it easier to implement. After this expansion, the number of nodes increases from 180 to 444, but this simplifies the network topology representation. They have a different functional topology of the network for each cell of parasegment since some proteins don't exist in that particular cell or on the neighboring cells. This boolean model is calculated again and again for different time points and different places. They used the patterns of segment polarity genes formed before stage 8 as initial states and the final stable state is wild type patterns maintained during stages 9-11. The modeling is done on only one parasegment, for which we can find all possible steady states of this network. They iterate the dynamical system defined by their Boolean model starting from the initial state described above. They found that after only six time steps, the expression pattern stabilizes in a time invariant spatial pattern.

They found 10 solutions of their model analytically which lead to six distinct steady states. Three of those steady states are well known experimentally, corresponding to wild type pattern and two mutant patterns with either no stripes or broadened stripes. The existence of three additional states suggests that the network can produce some patterns which are not needed for *Drosophila melanogaster* embryogenesis also. They also identified the basin of attraction for each steady state by searching in the space of potential initial conditions and observed that the network could correct itself for the errors in the initial expression patterns. This property of error correcting is a significant robustness property of the segment polarity network.

Their model gives several insights into the design of the segment polarity network. First of all, it suggests that the wingless gene is a key element in the network, and its initiation in the right pattern at the right time is vital. On the other hand, non-initiation of engrailed and hedgehog could be rescued by the interactions in the segment polarity network. In their study in order to reduce the number of assumptions taken they assumed that the interactions in the network follow a single time step. They extend this by using the assumption that decay of proteins is slower (twice of mRNA decay time). They make this change to the network to make it more realistic. The number of intermediate steps decrease for the two step model. This assumption provides a more realistic modeling of the decay of proteins without changing the conclusions of the model. The model of Albert & Othmer could be improved by taking more realistic assumptions such as taking different time intervals for the decay for mRNAs and proteins or considering a two dimensional array of cells instead of one.

As mentioned above the gene networks of embryonic segmentation in *Drosophila* have been modeled by differential equation models or Boolean models (von Dassow *et al*, 2000; Sanchez & Thieffry, 2001; Albert & Othmer 2003; Jaeger *et al*, 2004). The first modeling study which focused on the segment polarity gene network was done by von Dassow *et al* (2000) as described in the differential equation modeling section, where they developed a nonlinear differential equation model of the network. Their study focused on five genes ci, en, hh, ptc and wg and their proteins. Their choice of network topology failed to reproduce the wild type expression patterns of these genes, which they extended by adding two more interactions. This extension resulted in a robust network with respect to variations in the kinetic variables. The robustness of this network to

changes in variables, suggested that the topology of the network and the regulatory interactions within the network are essential for robustness. Based on this observation, Albert & Othmer (2003) used a Boolean model to reproduce the main characteristics of the segmentation polarity network. In their study they showed that the network topology and signatures of interactions in the network, whether an interaction is activating or inhibiting, is enough to reproduce the essential features of the segment polarity network dynamics. We should note here that although, Albert & Othmer (2003) modeled the segment polarity network with a simple Boolean model, other networks might require more comprehensive models which incorporate features such as asynchronous updating.

Although both studies come to the same conclusion, the networks they employed are slightly different. The difference between their choices of network topology is due to two opposing observations on *en* inhibition; Aza-Blanc *et al* (1997) suggests that *en* inhibition is due to CIR and Cadigan *et al* (1994) suggests that this inhibition might be due to transcription factors encoded by sloppy paired gene. Hence, in Albert & Othmer (2003) they added sloppy pair to get the asymmetrical *en* activation instead of taking *en* inhibition by CIR. On the other hand, the level of importance given to inhibition by these studies is different. In Von Dassow *et al* (2000) inhibitory effects only reduce the level of activation; however in Albert & Othmer (2003) inhibitory effects are dominant. This difference resulted in a large number of patterns with very broad en and wg stripes for even wild type initial gene expression patterns in von Dassow *et al* (2000).

Sanchez & Thieffry (2001) used a Boolean approach, similar to their earlier study on the dorso-ventral patterning (Sanchez *et al,* 1997) to model the gap gene network in *Drosophila melanogaster*. Up until now the experimental studies, especially mutation

analysis led the construction of a qualitative network, which determines the pattern formation in *Drosophila*. This study extends these analyses with an addition to the understanding of their dynamics and attractor states. Their model was able to simulate the wild type, single and multiple mutant qualitative gap gene expression patterns and describe the ways a gap gene regulatory network function to generate different patterns in response to maternal factors; Bcd, Hb and Cad. Their modeling also sheds light on the least number of functional levels associated with the maternal (Bch, Cad and Hb) and gap genes (Gt, Hb, Kr and Kni), the most crucial interactions and regulatory circuits of the gap gene network, the ordering of different regulatory interactions governed by each of these products according to corresponding concentration scales and the importance of gap-gap cross regulation in this network. For example, although cross-inhibition between gap genes was suggested as a critical mechanism for creating gap gene expression patterns, their network analysis suggested that cross inhibitory interactions between *gt* and *Kr* constitutes a positive circuit which is critical for the whole gap gene network, but not others (Rivera-Pomar & Jackle, 1996).

Their Boolean model takes the sum of the regulatory inputs to a target and transforms it into logical parameters. How this transformation is done is a critical part of the modeling. To select the specific values for parameters of the model, which takes care of the transformation, they dynamically analyzed the gap gene system and accepted the lowest parametric values that can generate the expression states compatible with known wild type and mutant phenotypes. These parameters lead to the finding of the most important cross regulatory interactions between the elements of the gap gene network. In their Boolean modeling they also divided the embryo to four domains along the anterior-

posterior axis, depending on the concentration levels of maternal morphogens (Bcd, Cad and Hb) and assigned the logical variables (maternal and gap genes) to different functional threshold levels. They assigned three functional thresholds for Bicoid (Bcd) and Hunchback (Hb), two for Caudal (Cad) and Kruppel (Kr), and one for Giant (Gt) and Knirps (Kni). In addition to that, if necessary, they ordered different functional interactions where distinct functional concentrations of the same regulatory product are involved. For example they assumed that Cad will activate kni at the first threshold and gt at the second threshold.

As discussed above in modeling dynamic biological processes differential equation models are used not to lose the fine details of the system. Those models use details of the system including production, diffusion and degradation rates of regulatory factors to develop a model that can typically be usually solved in a computationally expensive way.

However, in most of the biological systems we lack the good quality quantitative information on the molecular interactions between genes and proteins. Modelers in order to circumvent this problem use Boolean models instead of differential equation models. The Boolean model of Sanchez & Thieffry (2001) and the differential equation model of Jaeger *et al* (2004) on the gap gene network show two sides of the coin here. Although, Sanchez & Thieffry (2001) and Jaeger *et al* (2004) accomplish a comparable level of analysis for gap gene network, there are some key differences worth mentioning, which we will discuss further below.

A repressive feedback loop between kni and hb is reported in Jaeger *et al* (2004) as essential for the gap gene network, but not in Sanchez & Thieffry (2001)

due to the fact that they did not take into consideration *tll* and the posterior *hb* in their analysis. On the other hand Sanchez & Thieffry (2001) suggest a dual role for Hb in Kr regulation but Jaeger *et al* (2004) argues that the dual role of Hb is not required for the proper expression of Kr. However this might be due to the fact that, Jaeger *et al* (2004) takes the sum of the contributions from all the regulatory proteins which possibly excludes all potential context sensitive interactions. Jaeger *et al* (2004) suggests autoactivation as a critical component for sharpening gap domain boundaries, however Sanchez & Thieffry (2001) could not find it.

The differences in their results might also be due to the minimalist approach of the Boolean modeling. In logical analysis functional thresholds are assigned for continuous protein concentrations, which results in four discrete regions of functional borders for the gene expression domain along the anterior-posterior axis, however these borders are crude and perhaps do not match well with real expression borders. For this reason modeling boundary sharpening in detail similar to Jaeger *et al* (2004) is not possible for Sanchez & Thieffry (2001).

Another difference between these two approaches is the fact that the approach of Jaeger *et al* (2004) is computationally much more expensive than Sanchez & Thieffry (2001) despite the fact that it is applied for modeling a one dimensional strip of nuclei. On the other hand the success of Jaeger *et al* (2004)'s parameter estimation technique, which is the core part of their study, depends heavily on the very high quality of the data on gene expression levels. However this is not possible at this time.

Another application of a Boolean model is given by Yuh *et al* (2001), where the approach is used to understand the transcriptional regulation on the DNA level. In this paper they analyzed the *endo16* gene of sea urchin in detail, which encodes a protein of the embryonic and larval midgut. This enhancer is a relatively well studied i.e. regulatory elements which control this gene spatially and temporally are known. The regulatory organization of the endo16 gene has been created by many experimental studies which launched the regulatory regions, their functions and interrelations between each other. The *endo16* gene has a complex regulatory enhancer region, which helps it to react to diverse spatial and temporal conditions in development. There are not many developmentally regulated genes that have been analyzed in as much detail as *endo16*. Because of this detailed knowledge, the endo16 gene is usually used as an example to show how developmental enhancers process regulatory information.

A small module (module A) in the regulatory region of the *endo16* gene has been modeled by logic gates such as AND, OR, NOT to state the interactions between regulatory regions (Yuh *et al*, 1998). Recently they extended their earlier study by analysis of a nearby module (module B) and its interactions with module A. In this enhancer, module A functions like a central processing unit, regulating the expression in the early embryos and contributing to the expression in late embryos with module B. There is a relatively large gap(~240 bp) between module A and module B. Either piece of DNA, i.e., Module B or Module A, if associated with a reporter gene (they used Bp-CAT) and injected into eggs, independently displays a specific and characteristic transcriptional activity.

In their study they used mutational analysis to set the logical model. However using mutation data usually results in a recapitulation of what is known rather than new findings. They used quantitative measurements of the kinetic outputs of various embryonic expression constructs that had been introduced into fertilized eggs. In this way they could recognize regulatory functions. In their Boolean model they incorporated the repressive contributions from other modules (DC, E, and F) to module A, a module B to module A regulatory connection (nine different sequence specific transcription factors interact within Modules B and A), synergistic contributions of module A to module B and a control switch between module A and module B. Using their Boolean model they predicted that there is an internal switch in the endo16 enhancer region gene which moves the control from module A to module B. They confirmed their prediction experimentally and found the key players on module B which control this switch. In this enhancer, module A starts the activity of the gene at earlier time points in development (vegetal plate specification), however, once the gut differentiation starts module B takes control and becomes the primary operating unit. Module B generates and transmits its regulatory input to module A, and interacts with module A to amplify the expression of endo16. In Yuh *et al* (2001) they showed by a quantitative kinetic experiment that the expression of a construct including only module B was at all time points four to five times lower than a construct with module A and B together. Their model now could explain the control of expression changes in the endo16 gene throughout embryogenesis. Their model not only allowed them to summarize the interactions but also provided many testable predictions and predicted variations in the regulatory elements.

One shortcoming of Boolean models is their inherent deterministic behavior which may result in poor predictive power due to stochasticity in gene regulation and noise in the experimental measurements. The noise in the biological data makes application of Boolean models without probabilistic considerations impossible. One possible approach to solve this problem is adding stochastic functions to parts of the model where necessary.

Lack of knowledge in the network architecture is another problem for the use of Boolean models. Although, Boolean models could be used for investigative purposes, for biological systems with insufficient knowledge in system parameters, their success might be heavily affected by network architecture. It has been previously reported by the knock out experiments in the yeast network of Li *et al* (2004) that a single change in the network changes the dynamical trajectory with a 50% probability.

Sanchez & Thieffry (2001) and Yuh *et al* (2001) used analysis of mutation experiments to construct their models. However, this approach is simple, crude and likely to replicate the explanation of the data they are based on instead of suggesting new ways to explain it. Possibly if a new interaction is suggested by experiments in the system it will change the results of those types of studies. On the other hand the models, not based on assumptions coming from experimental data such as Jaeger *et al* (2004), have the chance to suggest new ways to explain the data.

The simplicity of Boolean models comes with a price; the accuracy in the system will be lost. For example, a Boolean variable has only two possible states, which is not a good simplification for many biological properties. If the system depends crucially on the

fine details of the system such as reaction rates, timing of regulatory relations and concentrations of mRNAs and proteins then these models will not do a satisfactory job at describing the system. For example if a gene is negatively regulating its own production, in Boolean modeling this would give a oscillatory behavior, but in reality such a process will lead to steady state unless there is a significant time delay.

PARAMETER ESTIMATION

Most biological modeling problems reduce to an inverse problem, where the certain parameters in the model should be estimated. However, inverse problems are not easy to solve, it is usually a challenge to find a unique parameter set which satisfies all the constraints of the system. On the other hand they are often computationally very expensive. Nevertheless, there is no universal parameter estimation technique which works well for all inverse problems; rather there are different parameter estimation methods that are customized for different problem types. The choice of a suitable parameter estimation technique and its validation is extremely important since this choice determines whether the problem is solved fast or slow, or even solved at all. However, this problem has not been treated in the biological modeling in particular transcriptional modeling literature explicitly (Janssens *et al*, 2006; Zinzen *et al*, 2006; Segal *et al,* 2008).

To estimate the parameters, which fit the model to experimental data, we must first identify an objective function. Objective function depends on variables of the model and gives a quantitative measure of the performance of the model. Often sum of squares of the residuals between the model's prediction and experimental data is employed as the objective function. The goal in parameter estimation is to find values of the parameters that optimize the objective. Often the parameters are restricted, or *constrained*, in some

way, usually expressed as equalities and inequalities. Parameter estimation algorithms start with an initial guess and iteratively generate new estimates until they stop, optimistically at a solution. How the iteration process works differentiate different parameter estimation techniques and usually depends on the objective function and constraints of the parameters. In biological models objective functions and the constraints are often nonlinear, which might imply multimodality i.e. the objective function might have many local and global optimums. The modelers are usually interested in finding the global optimum solution among the set of all possible solutions, however this is not an easy task.

Parameter fitting approaches could be divided into two main groups as local and global parameter estimation techniques. Local estimation techniques include the conjugate gradient method, Newton's method, simplex methods, and the Nelder-Mead method. Within all local techniques two groups are particularly important; gradient based approaches such as Newton's method and the Levenberg-Marquardt method and direct search methods such as the Nelder-Mead method. Gradient based approaches require the calculation of the objective function's derivative or at least its approximation by a finite difference method. On the other hand direct search methods circumvent the need for gradient calculation, which makes them the choice for problems with discontinuous, nondifferentiable or nonlinear objective functions. This method is based on a comparison of the objective function at the vertices of a simplex, which is updated at each step.

Use of local parameter estimation techniques for searching global optimums of biological models is troublesome; the search usually ends up in a local optimum rather than global (Mendes & Kell, 1998). When prior knowledge about the parameters is

available local techniques could be used, however for many biological problems this is not the case. To overcome this problem, local parameter estimation techniques might be used repeatedly with different initial starting points for parameters, however this approach is not very efficient. Mendes (2001) noted that gradient methods could not find the optimum parameter set from any random starting point for estimating 36 parameters of a nonlinear biochemical dynamic model.

Global parameter estimation techniques include deterministic strategies such as the branch and bound method, and interval optimization, and stochastic strategies such as genetic algorithms, the simulated annealing and, evolutionary strategies. Deterministic methods provide some guarantee of finding the global optimum, however finding the global optimum is computationally very expensive. On the other hand stochastic models are probabilistic approaches and give merely a weak guarantee on finding the global optimum. However they can reach the vicinity of the real solution in a reasonable amount of time.

Global parameter estimation techniques, especially stochastic strategies, are better than their local counterparts in finding the global optimum of the system and have been shown to be more suitable for biological systems (Mendes & Kell, 1998; Mendes 2001, Moles *et al*, 2003). In a recent study, Banga and colleagues compared a balanced selection of competitive parameter estimation algorithms, including several deterministic and stochastic global parameter estimation techniques, for their efficiency and reliability. They observed that the stochastic approach evolutionary strategy functioned best on their continuous problem with a high number of unknowns (Moles *et al*, 2003). Slow convergence is a problem for global methods, however many global methods have

parallel versions, which shorten the time needed to finish the job. In a recent study, Mendes & Kell (1998) considered the parameter estimation in the mechanism of irreversible inhibition of HIV proteinase, which has 20 parameters to estimate. They suggested that within all the methods they tried the simulated annealing algorithm functioned the best; however the solution obtained by the Levenberg-Marquardt was comparable and the algorithm was 750 fold faster. To overcome this problem, hybrid models, which combine global and local techniques, have been used frequently (Gursky *et al*, 2004).

In several recent studies, we have realized once more the necessity of a study which compares the performance of parameter estimation techniques on gene regulation models. In Reinitz & Sharp (1995), the lam simulated annealing method is used where each parameter estimation took approximately 1 week of CPU time on a Sparc 2 and in Jaeger *et al* (2004a, 2004b), a parallel lam simulated annealing method is used where each parameter estimation took between 8-160 hours on ten 2.4GHz Pentium P4 Xeon processors (Lam & Delosme, 1988). The same study is repeated with the same conclusions by more efficient algorithms. Perkins *et al* (2006) used local search procedures and particular characteristics of the gap gene system to estimate the parameters in 1-2 days. Fomekong-Nanfack *et al* (2007) used the evolution strategy approach (motivated by the earlier studies Moles *et al* (2003) and Runarsson & Yao (2000)) and reduced the time needed 5-140 times compared to the parallel lam simulated annealing method.

Parameter estimation techniques have been used in modeling extensively as mentioned above, however there are not many studies in the literature which compare

parameter estimation techniques in the field of modeling, particularly gene regulation modeling. Mendes & Kell (1998) discussed the performance of the parameter estimation techniques for finding global optimum in biochemical models and concluded that there is no best algorithm which works well for all problems. In their study they recommended the use of a set of diverse parameter estimation methods to attain the best possible solution.

Modeling approaches will get more complicated soon to incorporate the new genome scale data. To estimate the parameters of these models we need algorithms that are robust (starting point independent), efficient (not require much computational power and storage) and accurate (not sensitive to errors in the data). Although we cannot attain all of these goals at the same time, it is our job to pick the best for our needs.

MODEL SELECTION

Quantitative analysis of biological systems relies on the iterative process of integration of experimentation, data processing and modeling. New experiments and model development continues until an agreement is reached between the experimental data and model predictions. Modelers usually require plenty of good quality data for this iterative process; however the data collection from biological systems is usually limited, which makes the modeling extremely challenging. For example, in biological systems data is usually collected at a certain temporal state, i.e. it is only a snapshot of a dynamic system. Although there are some biological systems which are relatively better known, such as anterior-posterior patterning in *Drosophila*, even the current information in these systems is not sufficient to detail all the regulatory relations.

The models which are restricted by the limited amount of data could generally reproduce the experimental data, but not provide any new insights. Currently, there are widespread studies that generate genome, transcriptome and proteome data, which will end the data limitation problem. The new challenge would be how to merge these different "omic" data types. Recently several studies showed the power of integrating insights from different data sets such as transcriptome and genome sequence (Tavazoie *et al*, 1999; Segal *et al*, 2004). For example Tavazoie et al 1999 used gene expression data from microarrays to cluster co-expressed genes and searched upstream regions of these genes to identify common regulatory motifs. They applied this technique for identifying novel regulatory networks in *S. cerevisiae* and were able to identify 18 motifs in the upstream sequences of genes in 12 clusters that are overrepresented in their cluster and absent in the others. However, such a study is difficult to extend to many biological systems for now.

The noisy nature of biological data creates another challenge for modelers and affects the success of the modeling efforts dramatically. Quantitative biological data is usually noisy due to both inherent noisiness of the biological systems and the imprecision of the data collection methods. Although the noise can be reduced by repetition of the experiments and application of carefully chosen data processing steps, which estimate and remove the noise from the data, it is not possible to eliminate the noise from the biological data entirely. However, despite the noise, modeling such biological systems might still provide some critical insights which might not be obtained without modeling.

Understanding the nature of noise in biological data sets is critical for the selection of appropriate analysis instruments and modeling. The model chosen should be

able to deal with noisy data. As mentioned above some modeling problems reduce to an inverse problem i.e. identifying parameters of the model which fits the model to experimental data. However, the noisy data makes this inverse problem ill-posed where the solution lacks stability i.e. a few percent of noise can possibly lead to huge relative errors in the parameters and unreliable solutions. This problem might be overcome partly by approaches such as the Tikhonov regularization method.

A good model is consistent with the available data, reflects essential properties of the system, helps answer specific questions about the system such as computing properties difficult to measure, and guides the researcher to design new experiments. Before we start modeling we should have a good grasp of the system, we should know properties of the system such as major molecular components, their interactions, layers of interactions, and the physical geometry of the system. Modeling is a simplification of reality, and for this reason when we are modeling we should make a choice about the level of detail we want in the model. A highly detailed model that may give a more accurate representation of the system is not necessarily a good thing; the increase in number of parameters boosts the possibility of overfitting to the data. After the level of detail is decided, depending on the data available and questions we are interested in, we can choose a model type. A biological system can be investigated with different experimental methods and mathematical models. Although standardizing experimental conditions and modeling approaches are necessary for integration of efforts of different groups and comparability of results, it is hard to achieve. On the other hand, use of diverse modeling approaches will promote creativity and possibly decode different aspects of the problem. The choice of the model is largely dictated by the data available

and questions that are addressed. The model should be appropriate for the problem we are solving. All of the models mentioned above have both advantages and disadvantages. For example, a transcriptional network can be modeled by a discrete approach dynamic Boolean model, where the time is discrete, each node of the network (mRNA, protein or other component) has few states and the regulatory interactions between nodes are described by logical functions (Sanchez & Thieffry, 2001; Yuh *et al*, 2001). A transcription network can also be modeled by a continuous approach where each node of the network are continuous functions of time and the evolution of the nodes are modeled by differential equations, usually with mass action kinetics (von Dassow *et al*, 2000; Jaeger *et al*, 2004). Although differential equation models provide more detail than Boolean models the latter are simpler to handle, easier to expand to larger biological systems and are computationally efficient. However, Boolean models are time discrete and can only provide qualitative results. On the other hand, both of these approaches take the DNA level regulation as a black box and are not as useful as thermodynamic models for explaining the enhancer architecture.

The era of systems biology has brought together people from different disciplines for collaborative studies. For this reason, models proposed for biological problems should be easy to understand by researchers from other disciplines for its diverse applications. A model that is hard to code and implement may not be so useful since it is not easy for a researcher without a quantitative background to imitate or use those models and codes. However, a good model should not be computationally challenging and expensive. Although, the improvements in the computing make the latter problem less significant, it

is still a challenge. Although, these computational considerations don't make a model right or wrong; but rather affect its applicability and extendibility.

REFERENCES

Ackers GK, Johnson AD, Shea MA (1982) Quantitative model for gene regulation by λ phage repressor. *Proc Natl Acad Sci USA* **79(4):** 1129-1133

Albert R, Othmer HG (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. *J Theor Biol* **223(1):** 1-18

Aza-Blanc P, Ramírez-Weber FA, Laget MP, Schwartz C, Kornberg TB (1997) Proteolysis that is inhibited by hedgehog targets Cubitus interruptus protein to the nucleus and converts it to a repressor. *Cell* **89(7):** 1043-1053

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

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R (2005b) Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev* **15(2):** 125-135

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R (2005a) Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* **15(2):** 116-124

Buchler NE, Gerland U, Hwa T (2003) On schemes of combinatorial transcription logic. *Proc Natl Acad Sci USA* **100(9):** 5136-5141

Cadigan KM, Grossniklaus U, Gehring WJ (1994) Localized expression of sloppy paired protein maintains the polarity of Drosophila parasegments. *Genes Dev* **8:** 899–913

Coppey M, Berezhkovskii AM, Kim Y, Boettiger AN, Shvartsman SY (2007) Modeling the bicoid gradient: diffusion and reversible nuclear trapping of a stable protein. *Dev Biol* **312(2):** 623-630

Crocker J, Tamori Y, Erives A (2008) Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol* **6:** e263

de Jong H (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* **9(1):** 67-103

Eldar A, Dorfman R, Weiss D, Ashe H, Shilo BZ, Barkai N (2002) Robustness of the BMP morphogen gradient in Drosophila embryonic patterning. *Nature* **419(6904):** 304-308

Fakhouri WD, Ay A, Sayal R, Dresch J, Dayringer E, Chiu C, Arnosti DN Deciphering a transcriptional grammar: modeling short-range repression in the *Drosophila* embryo (in preparation).

Fomekong-Nanfack Y, Kaandorp JA, Blom J (2007) Efficient parameter estimation for spatio-temporal models of pattern formation: case study of Drosophila melanogaster. *Bioinformatics* **23(24):** 3356-3363

Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data *J Comput Biol* **7:** 601-620

Gertz J, Cohen BA (2009) Environment-specific combinatorial cis-regulation in synthetic promoters. *Mol Syst Biol* **5:** 244

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* **81:** 2340–2361

Goodwin B (1965) Oscillatory behaviour in enzymatic control process. *Adv Enz Regul* **3:** 425-438

Granek JA, Clarke ND (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol* **6(10):** R87

Gregor T, Bialek W, de Ruyter van Steveninck RR, Tank DW, Wieschaus EF (2005) Diffusion and scaling during early embryonic pattern formation. *Proc Natl Acad Sci USA* **102(51):** 18403-18407

Gregor T, Wieschaus EF, McGregor AP, Bialek W, Tank DW (2007) Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell* **130(1):** 141-152

Griffith JS (1968) Mathematics of cellular control processes. II. Positive feedback to one gene. *J Theor Biol* **20**: 209-216

Gursky V.V., Jaeger J., Kozlov K.N., Reinitz J., Samsonov A.M. (2004). Pattern formation and nuclear divisions are uncoupled in Drosophila segmentation: Comparison of spatially discrete and continuous models. *Physica D*, **193**: 286-302.

Ip YT, Park RE, Kosman D, Bier E, Levine M (1992) The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the Drosophila embryo. *Genes Dev* **6:** 1728-1739

Jaeger J, Blagov M, Kosman D, Kozlov KN, Manu, Myasnikova E, Surkova S, Vanario-Alonso CE, Samsonova M, Sharp DH, Reinitz J (2004) Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster. *Genetics* **167(4):** 1721-1737

Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, Kozlov KN, Manu, Myasnikova E, Vanario-Alonso CE, Samsonova M, Sharp DH, Reinitz J (2004) Dynamic control of positional information in the early Drosophila blastoderm. *Nature* **430(6997):** 368-371

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

Keränen SV, Fowlkes CC, Luengo Hendriks CL, Sudar D, Knowles DW, Malik J, Biggin MD (2006) Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution II: dynamics. *Genome Biol* **7(12):** R124

Lam J, Delosme J (1988) Performance of a New Annealing Schedule," *DAC*, 306 -311

Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. *Proc Natl Acad Sci USA* **101:** 4781-4786

Ludwig MZ, Kreitman M (1995) Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. *Mol Biol Evol* **12:** 1002-1011

Luengo Hendriks CL, Keranen SVE, Fowlkes CC, Simirenko L, Weber GH, DePace AH, Henriquez C, Kaszuba DW, Hamann B, Eisen MB, Malik J, Sudar D, Biggin MD, and Knowles DW (2006) Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution 1: data acquisition pipeline. *Genome Biol* **7:** R123

Mackey MC, Santillán M, Yildirim N (2004) Modeling operon dynamics: the tryptophan and lactose operons as paradigms. *C R Biol* **327(3):** 211-224

Mendes P, Kell D (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* **14(10):** 869-883

Mendes, P. and Kell, D.B. (2001) MEG (Model Extender for Gepasi): a program for the modelling of complex, heterogeneous, cellular systems. *Bioinformatics*, **17,** 288-289.

Moles CG, Mendes P, Banga JR (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res* **13(11):** 2467-2474

Nicolis G, Prigogine I (1977) *Self-organization in nonequilibrium systems. From dissipative structures to order through fluctuations.* Wiley, New York, USA

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

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in C.* Cambridge University Press, Cambridge, UK

Reinitz J, Sharp DH (1995) Mechanism of eve stripe formation. *Mech Dev* **49(1-2):** 133-58

Rivera-Pomar R, Jackle H (1996) From gradients to stripes in *Drosophila* embryogenesis: filling in the Gaps. *Trends Genet* **12:** 478-483

Runarsson TP, Yao X (2000) Stochastic Ranking for Constrained Evolutionary Optimization *IEEE Transactions on Evolutionary Computation* **4(3):** 274-283

Sánchez L, Thieffry D (2001) A logical analysis of the Drosophila gap-gene system. *J Theor Biol* **211(2):** 115-141

Sanchez L, van Helden J, Thieffry D (1997) Establishement of the dorso-ventral pattern during embryonic development of drosophila melanogasater: a logical analysis *J Theor Biol* **189(4):** 377-389

Santillan M, Mackey MC (2001) Dynamic behavior in mathematical models of the tryptophan operon. *Chaos* **11(1):** 261-268

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

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34(2):** 166-176

Shea MA, Ackers GK (1985) The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol* **181(2):** 211–230

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

Szymanski P, Levine M (1995) Multiple modes of dorsal-bHLH transcriptional synergy in the Drosophila embryo.*EMBO J* **14(10):** 2229-2238.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* **22(3):** 281-285

Turing AM (1952) The chemical basis of morphogenesis. *Philos Trans R Soc Lond B* **237:** 37-72

Vilar JM, Leibler S (2003) DNA looping and physical constraints on transcription regulation. *J Mol Biol* **331(5):** 981-989

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

Von Hippel PH, Revzin A, Gross CA, Wang AC (1974) Non-specific DNA binding of genome regulating proteins as a biological control mechanism: 1. The lac operon: equilibrium aspects. *Proc Natl Acad Sci USA* **71(12):** 4808-4812

Yildirim N, Mackey MC (2003) Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data *Biophys J* **84:** 2841-2851

Yuh CH, Bolouri H, Davidson EH (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279(5358):** 1896-1902

Yuh CH, Bolouri H, Davidson EH (2001) Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* **128(5):** 617-629

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

Chapter II

Image Processing and Analysis for Quantifying Gene Expression from Early *Drosophila* Embryos

Ahmet Ay, Walid D. Fakhouri, Chichia Chiu and David N. Arnosti, (2008). Image Processing and Analysis for Quantifying Gene Expression from Early *Drosophila* Embryos. *Tissue Engineering Part A.* **14(9):** 1517-1526.

ABSTRACT

Correlation of quantities of transcriptional activators and repressors with the mRNA output of target genes is a central issue for modeling gene regulation. In multicellular organisms, both spatial and temporal differences in gene expression must be taken into account; this can be achieved by use of in situ hybridization followed by Confocal Laser Scanning Microscopy (CLSM). Here we present a method to correlate the protein levels of the short-range repressor Giant with *lacZ* mRNA produced by reporter genes using images of *Drosophila* blastoderm embryos taken by CLSM. The image stacks from CLSM are processed using a semi-automatic algorithm to produce correlations between the repressor levels and *lacZ* mRNA reporter genes. We show that signals derived from CLSM are proportional to actual mRNA levels. Our analysis reveals that a suggested parabolic form of the background fluorescence in confocal images of early *Drosophila* embryos is evident most prominently in flattened specimens, with intact embryos exhibiting a more linear background. The data extraction described in this paper is primarily conceived for analysis of synthetic reporter genes that are designed to decipher *cis*-regulatory grammar, but the techniques are generalizable for quantitative analysis of other engineered or endogenous genes in embryos.

INTRODUCTION

Complex patterns of gene expression underlie the development of multicellular organisms, and capturing the spatial and temporal information of gene expression is critical for modeling gene regulatory networks. Microarrays or qPCR reactions can provide quantitative information, but these methods usually lack spatial information. On

the other hand, fluorescent in situ hybridizations provide spatial information about gene expression and have the possibility of providing quantitative information as well. Recent studies have relied on such in situ techniques to quantitate the levels of nuclear proteins in the *Drosophila* embryo for the purposes of modeling.[1,2]

Because of its extensively researched genetic network, the *Drosophila* embryo provides one of the best characterized systems for modeling transcriptional regulation. Transcription factors encoded by the maternal, gap and pair rule genes form a regulatory network, whose interactions have been carefully described in molecular studies.[3] A comprehensive mathematical description of this system still eludes us, however, in recent years several studies have modeled parts of this system.[1,2,4,5] In most cases, confocal images of early *Drosophila* embryos were used to provide data on levels of transcription factors in the nucleus. One study reported a straightforward approach in which levels of regulatory proteins in the nucleus were related to protein levels of downstream targets, while other studies have focused on quantitative descriptions of mRNA levels in the embryo.[6-8] A later study took a further step in correlating nuclear transcription factor levels to mRNA levels of a target gene but did not fully explore parameters and methods required for this analysis.[4] Here, we report such a method that correlates the level of transcription factor to level of reporter gene mRNA, as a basis for mathematical modeling of gene regulatory elements.

Key to quantitative assessment of gene expression levels is information about the proportionality of signal read out to the actual levels of mRNA and protein. This relationship has been insufficiently tested; therefore we examine this issue using gene dosage studies and independent mRNA measurements. Background (in this case,

nonspecific fluorescence) is another central issue in many biological data analyses. A simple approach is to apply uniform background subtraction from the data. It has been previously noted that the background for fluorescently stained *Drosophila* embryos can be represented as a paraboloid function.[9] Our study suggests that fluorescence background from undistorted embryos is not very paraboloidal, but as samples are flattened a paraboloidal background becomes more evident.

This methodology appears to provide the correct basis for quantitative modeling approaches that utilize empirically established gene regulatory surfaces to facilitate parameter fitting. Such quantitative approaches will provide the tools to discover important regulatory information in genomic data sets, as well as lay a foundation for design of engineered transcriptional elements.

MATERIALS AND METHODS

1. Immunofluorescent in situ hybridization:

Embryos were collected and fixed as previously described.[10] Immunofluorescent in situ hybridization was done essentially as previously described with some modifications.[11,15] All washes were done in 1.0 ml. Fixed embryos stored at –20°C in methanol were briefly washed six times with 100% ethanol, then with xylene for 1.0h. About 50µl of embryos were transferred into individual microfuge tubes, washed four times with methanol-phosphate buffer 0.1%-Tween80 ((PBT; 1.37M NaCl, 43 mM Na_2HPO_4, 14mM NaH_2PO_4),1:1, v/v ratio) and then with PBT four times, each for 3 min with continuous rocking. Embryos were washed in PBT with hybridization solution (50% formamide, 5X SSC (3M NaCl, 0.3M Na-citrate), 100µg/ml sonicated salmon sperm

DNA, 50μg/ml heparin, 0.1% Tween80 (1:1, v/v ratio) for 10 min, and then briefly in 100% hybridization solution for 2 min. New hybridization solution was added and the tubes were placed for 1h in a water bath at 55 °C. Anti-sense RNA probes of digU labeled *lacZ* were heated in 50 μl hybridization solution at 80°C for 3 min, directly placed on ice for 1 min, then prehybridization solution was completely removed and the probe in 50μl of hybridization solution was added to each tube, and tubes were incubated at 55°C for 18-20h. After incubation, 1ml of 55 °C hybridization solution was added to each tube and all tubes were rocked at room temperature for 1min, hybridization solution was changed and tubes were incubated for another 1h at 55 °C, followed by four washes with hybridization solution for 15 min each at 55 °C and with hybridization solution and PBT (1:1, v/v ratio) two times at room temperature for 15 min. Five more washes were done with PBT for 10 min with rocking at room temperature. The embryos were washed with a blocking solution 1.0% casein in maleic acid buffer (Western Blocking Reagent, Roche, Indianapolis, IN) with PBT (1:1, v/v ratio). 0.5 ml PBT and blocking solution (1:1, v/v) containing primary antibodies (2.2 μl of 1:250 dilution of mouse anti-DigU (Roche, Indianapolis, IN) and 0.75 μl of 1:800 dilution of rabbit anti-Giant (antibodies are a gift from Reinitz Lab[12]) was added and the tubes were rocked at 4 °C overnight. Tubes were washed four times each with PBT for 15 min at room temperature. 0.4 ml of PBT and 0.5 % casein blocking reagent (1:1) containing 8.0 μl of secondary antibodies: goat anti-mouse conjugated to Alexa 555 for detection of *lacZ* mRNA and chicken anti-rabbit conjugated to Alexa 488 for detection of the Giant protein (Molecular Probes, Eugene, OR) preabsorbed against fixed *yw* embryos were added to each vial, and the tubes were covered with aluminum foil to protect them from light and incubated overnight at 4 °C.

Embryos were then washed with PBT four times at room temperature for 5 min with rocking, washed in glycerol + PBT (7:3, v:v ratio) for 2h until the embryos settled to the bottom of the tubes. The embryos were then resuspended in 0.4 ml glycerol + PBT (9:1 ratio) and 0.2 ml of Permafluor Mountant Medium 434990 (Thermo Electron Corporation, Pittsburgh, PA), mounted labeled slides and covered with large rectangular Corning cover slips (No. 1.5, 24x50mm). The slides were protected from light and stored flat at room temperature until they were imaged.

Sequences for Reporter genes:

The reporter constructs analyzed here are synthetic, Twist and Dorsal activator binding sites were previously characterized in *lacZ* reporter genes in the embryo. [13] High-affinity Giant binding sites derived from the Kruppel promoter were previously characterized. [14] The 25bp neutral spacer used lacks high affinity sites to blastoderm transcriptional regulatory proteins. Two reporters are tested here, both containing a core of two Twist sites 5' of two Dorsal sites. A pair of Giant motifs is located either 31bp or 131bp 5' of the activators. Binding sites are capitalized; Giant sites are italicized, Twist sites are bold and Dorsal sites are underlined.

2gt.25.2T2D

5'-gaattc*TATGACGCAAGA*atgcgactcg*TATGACGCAAGA*ggatctggttagtaagctgtaaactggatc

c**CATATG**ttgag**CATATG**tctagaGGGATTTTCCCAaatcgaGGGAAAACCCAAccgcg

g-3'

2gt.125.2T2D

5'-gaattc*TATGACGCAAGA*atgcgactcg*TATGACGCAAGA*ggatctggttagtaagctgtaaactggatct

ggttagtaagctgtaaactggatctggttagtaagctgtaaactggatctggttagtaagctgtaaactggatctggttagtaagctg

taaactggatcc**CATATG**ttgag**CATATG**tctagaGGGATTTTCCCAaatcgaGGGAAAACCC

AAccgcgg –3'

2. Confocal Laser Scanning Microscopy (CLSM):

An inverted CLSM (Olympus, Flowview FV1000) was used for capturing the confocal fluorescent images. For each scan of mounted embryos, the same microscope settings were employed to all images to simplify comparison of results. The argon laser (488 nm) was set at 5.0% and helium-neon laser (543 nm) was set at 25%. Emitted fluorescence from Alexa 488 and 555 was detected through a dichroic 405/488/543 and a BP505-525 filter for the green channel and a BP560-620 filter for the red channel. The pinhole was set to 105 μm (1.0 Airy unit), and the PMT detector, gain and offset were 680, 1.0% and 6% for both green and red channels. The PMT detector was adjusted in cases where the embryos showed saturation of signal intensities. Embryos were imaged at a scan speed 6.51 s/scan, line filter equal 2 line Kalman filter, and 1.73 μm-thick Z-sections for a total of 21-30 slices. CLSM image data were stored as two separate stacks and projections of images for each channel. The section dimensions were 333μm in length and width and 1.73 μm in depth. Fluorescence pixels were recorded as 12-bit images and stored as Tiff files.

3. Determining proportionality of fluorescence intensity:

To test whether immunofluorescent signals were proportional to the actual levels of protein or mRNA, we varied gene dosage of *lacZ* reporter or the Giant repressor, assuming that heterozygotes express at half the level of homozygotes. Male transgenic

flies homozygous for a *lacZ* reporter gene regulated by the Giant repressor and Twist and Dorsal activators were crossed with virgin *yellow white* females to produce heterozygous embryos carrying a single copy of the *lacZ* gene. Homozygous *lacZ* reporter embryos were also collected. Two to four hour old homozygous and heterozygous embryos were dechorionated, fixed and analyzed for *lacZ* expression using immunofluorescent in situ hybridization. For RTqPCR mRNA analysis, dechorionated embryos were frozen in liquid nitrogen and stored at ⁻80 °C until needed. For extraction of total RNA, frozen embryos were dipped briefly in liquid nitrogen and macerated with a pestle in 500 µl of TriaZol reagent from Invitrogen (Carlsbad, CA, USA) in a 1.5 ml microfuge tube. The extraction and subsequent DNase (Roche, USA) treatment were performed according to the manufacturer's directions. To test for the presence of any DNA contamination, standard PCR was performed using primers specific for *lacZ* gene. RNA was spectrophotometrically quantified and equivalent amounts of RNA from each sample were used for the reverse transcriptase reaction using the Superscript III enzyme. The RTqPCR specific primer set for *lacZ* (F: 5'-CTGGGATCTGCCATTGTCAGA; R: 5'-TGGTGTGGGCCATAATTCAATT) was designed using Primer Express Software (ABI 7500 Prism). RTqPCR normalization and analysis were done as previously described previously.[15,16] Primer sets of actin (F: 5'-CGCGGTTACTCTTTCACCA; R: 5'-GCCATCTCCTGCTCAAAGTC) and 28S rRNA (F: 5'-GATGCCGCGCTAGTTACAT; R: 5'-GCTGCTCAACCACTTACAACAC) were used for normalization.

For analysis of Giant protein levels, a gt^{X11} stock was obtained from the Bloomington Stock Center.[17] gt^{X11} virgins were out-crossed with wild type male flies to

obtain male hemizygous gt^{X11} mutant embryos, gt^{X11} and heterozygous female embryos and wild-type female embryos.

4. Image processing prior to *lacZ* mRNA and Giant repressor measurement:

In this study, Image J (rbs.info.nih.gov) and MATLAB softwares (MathWorks) were used. ImageJ software was used to extract the fluorescent intensity levels of *lacZ* and Giant from each confocal fluorescent image. Software implementing the algorithms described below was written in MATLAB and is available upon request. Approximately 18-21 non-overlapping 1.73 μm depth optical sections were taken from each embryo, providing a 1024x1024 pixel 12 bit image representing the top half of the embryo. The slices were projected by taking maximum intensity for each pixel along the z-axis and the images were converted to 8 bit images. For flatter embryos analyzed in a previous study, the average of two adjacent sections was used but with our more three dimensional imaging, this approach is not valid because for most sections a portion of the central part of the embryo containing only yolk is included.[6]

To position images in a uniform horizontal manner, masks were used to rotate the images as described in Janssens *et al.*[6]. A mask is created for the projected images by using the threshold value that is taken from outside of the embryo. The convex hull of the embryo boundary is taken, the inside of the convex hull is set to 255, and the outside is set to 0. The rotated images were checked to compare the alignments of the two rotated channels. Images were flipped upside-down or left-right to align uniformly on the anterior-posterior and dorsal-ventral axes. To remove extraneous portions of the image outside the embryo, the rotated images were cropped to a minimal canvas size. The images were not changed to a uniform size in order to avoid interpolation of the data,

which might have varying effects for different embryos. Subsequent to these steps, perimeter pixels of the embryo were found using the mask.

In the next step, embryo boundaries, *lacZ* mRNA data and Giant repressor protein data were used to decide where to set boxes that are approximately the size of an average "cell". The embryos are at this point incompletely cellularized, but separate nuclei are clearly discernible. Positioning of the boxes is explained in data collection section. The Giant protein and *lacZ* mRNA levels were averaged inside those boxes and plotted in 2-dimensional space.

5. Comparison of imaging to data from other databases:

A variety of methods have been applied to obtain spatially resolved mRNA and protein information from *Drosophila* embryos. Our CLSM images of Giant were obtained using antibody staining and one photon imaging, but embryos were not flattened as in a previous study.[1] Alternatively, mRNA quantitation from non deformed embryos has been acquired by two photon imaging.[8] We compared qualitative features of our Giant images with these data sets to see if general trends were consistent. As an example, we observe in some of the images lower apparent levels of Giant protein in the ventral and dorsal regions of the embryo. Giant images were accessed from two other databases: the Berkeley *Drosophila* Transcription Network Project (BDTNP) and the Database of Segmentation Gene Expression in *Drosophila* (FlyEx). mRNA levels in actual embryos from the BDTNP and protein levels from an average virtual embryo image from the FlyEx were used.[8,18] Overall relative differences in intensities between anterior and posterior levels of Giant protein and mRNA were observed in all three data sets, as well as the relative lower levels of Giant in ventral regions of the anterior stripe (data not

shown). Despite differences in imaging, these methods appear to capture the same essential features of this gene's expression.

RESULTS

Background:

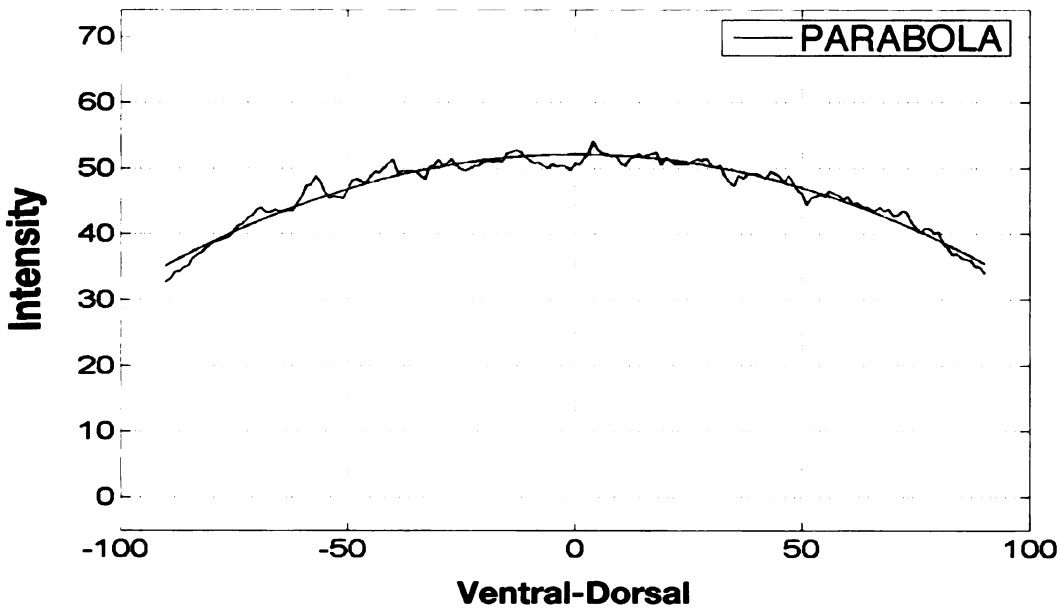
A critical issue in quantitative measurement of gene expression is background subtraction. In a previous study, it was noted that the background coming from nonspecific binding of a variety of primary and secondary antibodies to *Drosophila* embryos can be approximated by a paraboloid.[9] In that study, embryos were flattened (by the weight of a coverslip and using reduced amounts of mounting solution) so that a significant fraction of nuclei of the embryo could be captured in two 2μm sections. In our system, we do not distort the embryo, thus we tested whether this parabolic background relationship still applied. By assessing fluorescence in two different channels in embryos lacking a *lacZ* reporter gene, as well as in portions of embryos devoid of Giant protein (45-55% egg length) we found that the background often did not show a paraboloid shape (Fig. II-1A, B). One evident difference between our methods and that described in Myasnikova *et al.*[9] is the degree of flatness of the embryos. We imaged embryos of successively flatter proportions and analyzed background levels as a function of flatness (Fig. II-1C-F). Embryo flatness was judged by the number of 1.73μm slices required to reach the center of the embryo. Embryos that were flattened by a weighted coverslip had radial thicknesses ranging from 14-23μm, while rounder embryos had thicknesses of 24-33μm. This level of flattening is not as pronounced as that described in Myasnikova *et al.*[9] where almost the entire upper half of the embryo can be scanned in two 2μm slices, but a clear trend emerged. The background has a distinct tendency to be more

paraboloidal with flatter embryos, perhaps because unique light scattering properties of flat embryo sections contribute to a nonlinear background. Correlation between flatness of the embryos and curvature of the background intensity is measured by Pearson correlation coefficients. For *lacZ* images, the Pearson correlation coefficients between flatness of the embryo and curvature of the background curve are -0.5 ($p=0.006$) (ventral-dorsal) and -0.6 ($p<0.001$) (anterior-posterior). When this relationship is measured as the correlation between embryo flatness and the natural log of the curvature, the correlation coefficients are -0.6 ($p<0.001$) (ventral-dorsal) and -0.7 ($p<0.001$) (anterior-posterior). The relationship between flatness of the embryo and curvature of the background curve can be explained better with nonlinear functions. The main point here is that parabolic background cannot be assumed without taking the geometry of embryo into account. We also noted that for some embryos, inhomogeneously distributed background was evident in both channels (488 nm and 555 nm), suggesting that specific structural features of embryos can affect background (data not shown).

Figure II-1: Parabolic and nonparabolic background forms.

Representative parabolic (A) and flat (B) backgrounds. Data was taken from the middle 45-55% egg length of embryos . In general, we observed parabolic background for flatter embryos. Embryo cross sectional radii were measured and classified as 'flattened' (14-23 μm) or 'round' (24-33 μm). Considerable variability in curvature was noted, but flatter embryos tended to exhibit higher curvature overall. Data were obtained for Giant imaging ventral-dorsal (C), anterior-posterior (D) and *lacZ* imaging ventral-dorsal (E), anterior-posterior (F). Parabolicity of background is measured by curvature of the parabola fit to curve, with apex generally at center of embryo cross section. For anterior-posterior Giant background measurements, young gt^{XII} mutants expressing virtually no detectable Giant protein were used. For *lacZ* background imaging embryos without the reporter gene were utilized.

A.


PARABOLA
70
60
50
40
30
20
10
0
Intensity
-100
-50
0
50
100
Ventral-Dorsal

B.

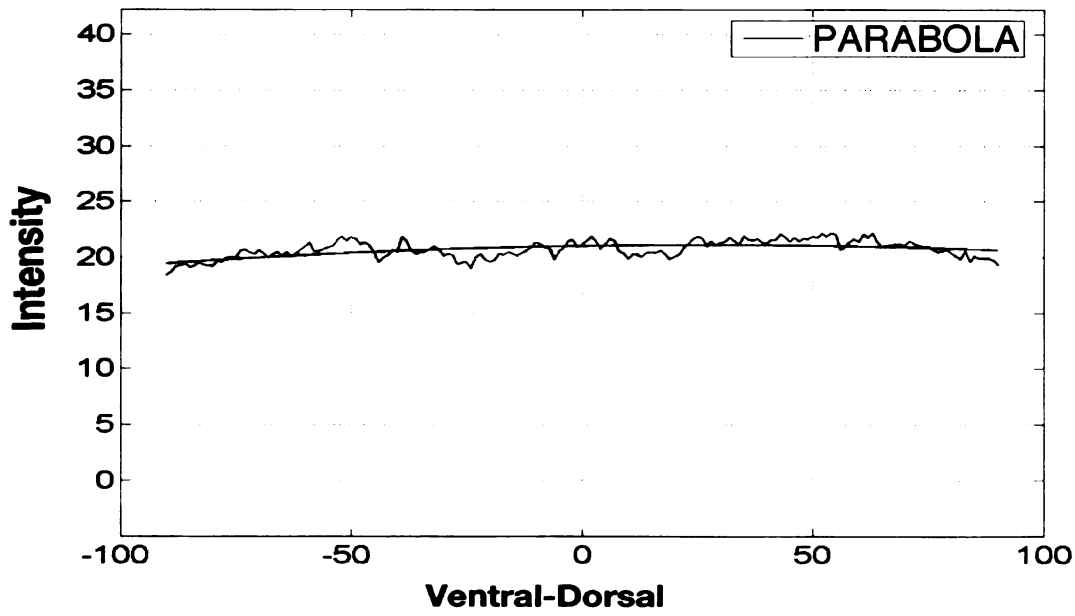
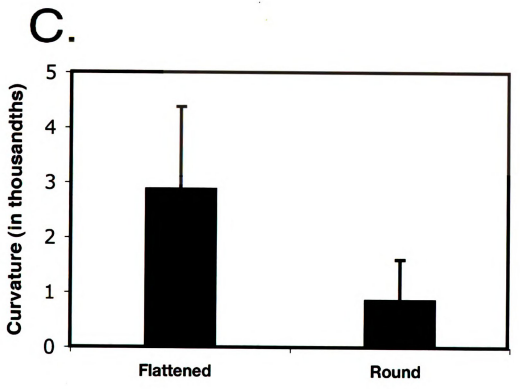

PARABOLA
40
35
30
25
20
15
10
5
0
Intensity
-100
-50
0
50
100
Ventral-Dorsal

Figure II-1: Continued.



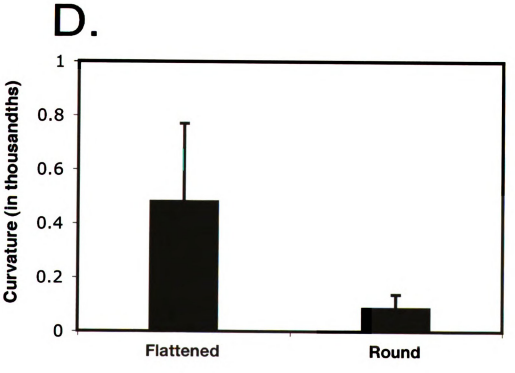
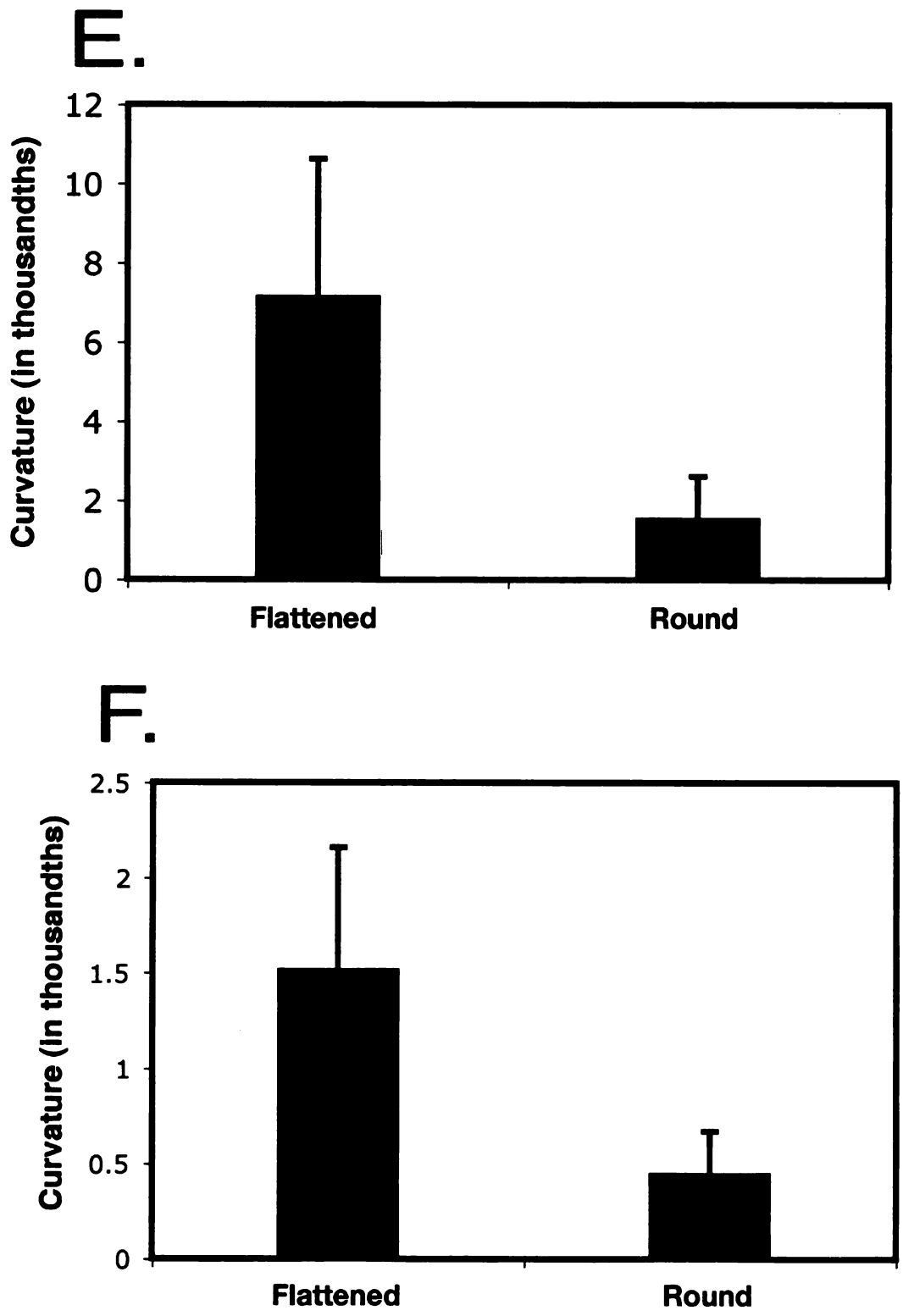


Figure II-1: Continued.



Fluorescent quantitation of mRNA and protein:

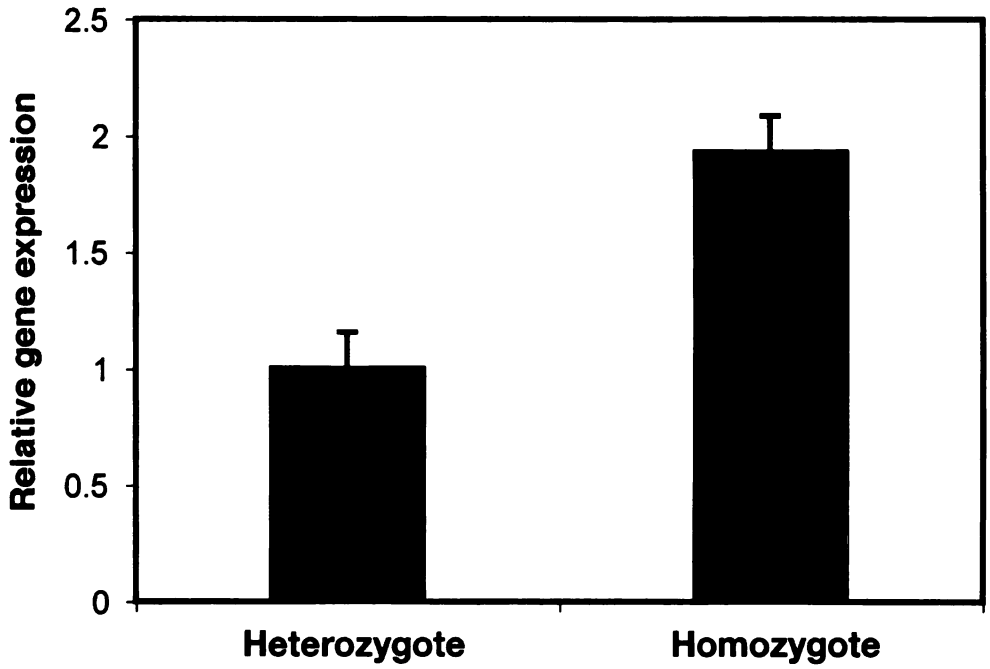
Quantitative measurements of transcription factor levels and mRNA inputs are essential for modeling gene regulation, yet surprisingly few studies relying on CLSM have independently tested whether the signals thus obtained exhibit strongly nonlinear effects. A quantitative study of *Drosophila* mRNA levels made a single correlation between *knirps* mRNA and Knirps protein obtained by CLSM as a test for proportionality, while other studies have quantitated levels of GFP protein as a proxy for transcript levels per cell.[8,19] We concluded that more rigorous independent measurements are essential for testing the validity of our quantitation. In the first case of mRNA detection, we varied the gene dose of a *lacZ* reporter expressed in ventral regions of the embryo and compared levels of mRNA by RTqPCR and CLSM. The RTqPCR results showed that the relative amount of *lacZ* mRNA was 1.0 in heterozygotes compared to 1.92 in homozygotes (Fig. II-2A). In comparison, fluorescent intensities for *lacZ* heterozygous compared to homozygous embryos were 1 to 1.85 (Fig. II-2B). This result suggests that both methods are responding to estimated twofold differences in mRNA in a similar manner, and that the fluorescent intensities do provide a reasonable proxy for actual mRNA levels.

Regarding protein measurement, previous studies have analyzed relative expression levels of a number of regulatory proteins in the *Drosophila* embryo by CLSM, but to our knowledge the proportionality of these measurements to actual protein has not been assessed.[1] We tested the measurement of protein by comparing the values of Giant expression in embryos derived from an outcross of gt^{XII} heterozygotes to a wild type strain (Fig. II-3). *gt* is located on chromosome 1, thus male hemizygotes carrying the

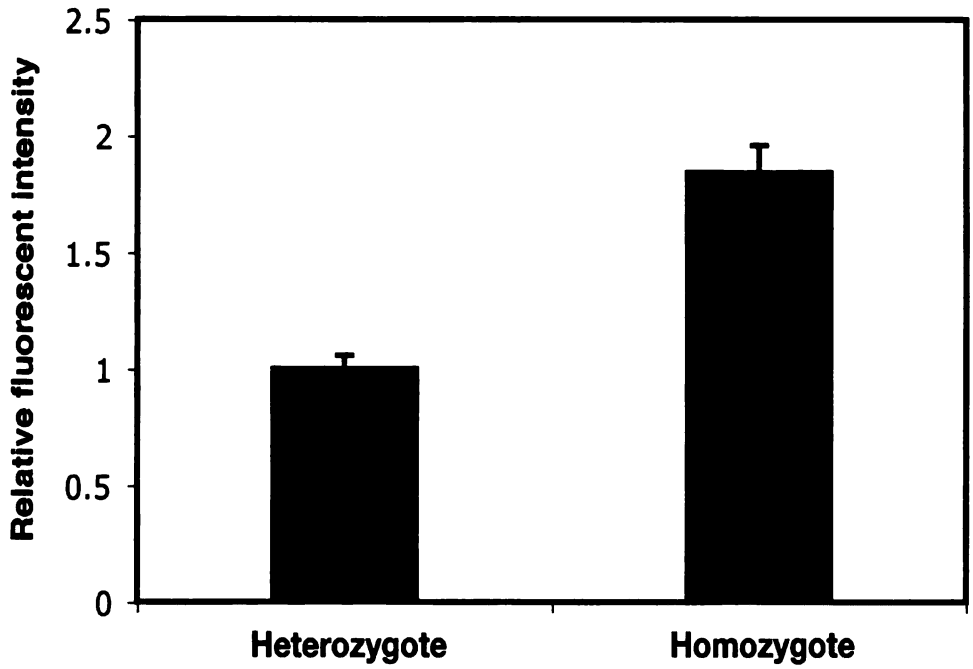
Figure II-2: Proportionality of signal to mRNA levels

The average relative amounts of *lacZ* mRNA in heterozygous and homozygous transgenic embryos were measured by RTqPCR analysis (A) and by immunofluorescent in situ hybridization and confocal laser scanning microscopy (B). Quantitation of mRNA levels by RTqPCR is representative of two biological assays; error bars indicate standard deviation from five technical replicates. Fluorescent imaging of 39 embryos were quantitated in (B).

A.


2.5
2
1.5
1
0.5
0
Relative gene expression
Heterozygote
Homozygote

B.


2.5
2
1.5
1
0.5
0
Relative fluorescent intensity
Heterozygote
Homozygote

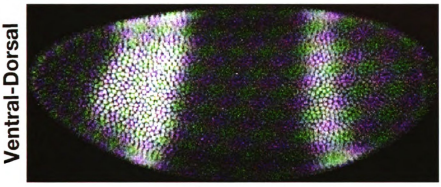
mutant allele would be expected to express the lowest levels of the protein, and female heterozygotes would have intermediate levels. The allele used, gt^{X11}, has been reported to express low levels of protein and consistent with this observation, no mid-blastoderm embryos entirely lacked Giant expression.

Embryos were stage matched to allow direct comparison of protein levels. Embryos with apparently normal intensity of Giant signal were observed alongside of age-matched embryos with very low Giant levels (Fig. II-3A, B); those differences are not likely to be merely a function of overall staining efficiency because background levels were fairly constant and such large differences are not usually seen when imaging wild-type embryos. When background subtracted, normalized levels of Giant were plotted for a set of 22 images, a roughly ten-fold range of signal intensities were observed (Fig. II-3C). The values did not fall into three obviously discrete clusters, corresponding to hemizygous null, heterozygous, and wild-type backgrounds, but were rather continuously distributed with some degree of over representation at higher and lower values. This effect may be due to nonlinearity of the fluorescent readout, which would produce some compression at one end of the spectrum, variability in detection or even expression of Giant from embryo to embryo, or some combination of these effects. We can conclude that this method does permit an apparent dynamic detection range of at least ten fold sufficient to capture the total dynamic variation reported for Giant in the blastoderm embryo but clearly a more detailed comparison of signal proportionality is warranted.[1] In light of our mRNA results, it appears that fluorescent detection can be an appropriate proxy for in situ levels of biomolecules, but the proportionality of read out must be established for each set of reagents.

Figure II-3: Proportionality of signal to Giant protein levels

Embryos with apparently normal intensity of Giant (A) and low intensity of Giant (B) were observed for age-matched embryos. Normalized levels of 22 background-subtracted Giant images were plotted, and a roughly ten-fold range of signal intensities were observed (C). Overlapping data points were separated into two columns for clarity for the anterior and posterior stripe.

A.


Ventral-Dorsal
Anterior-Posterior

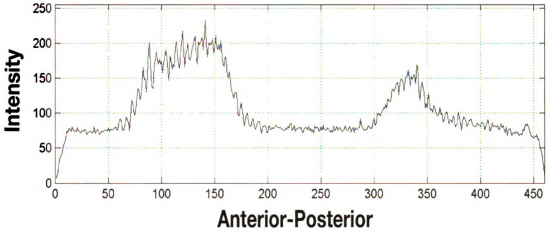

Intensity
250
200
150
100
50
0
0
50
100
150
200
250
300
350
400
450
Anterior-Posterior

Figure II-3: Continued.

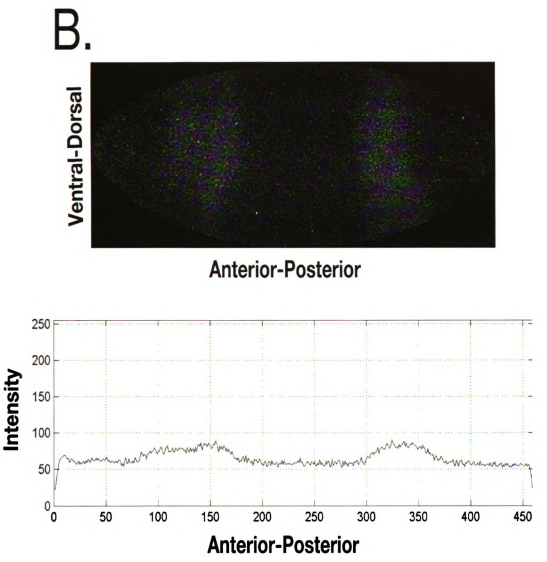
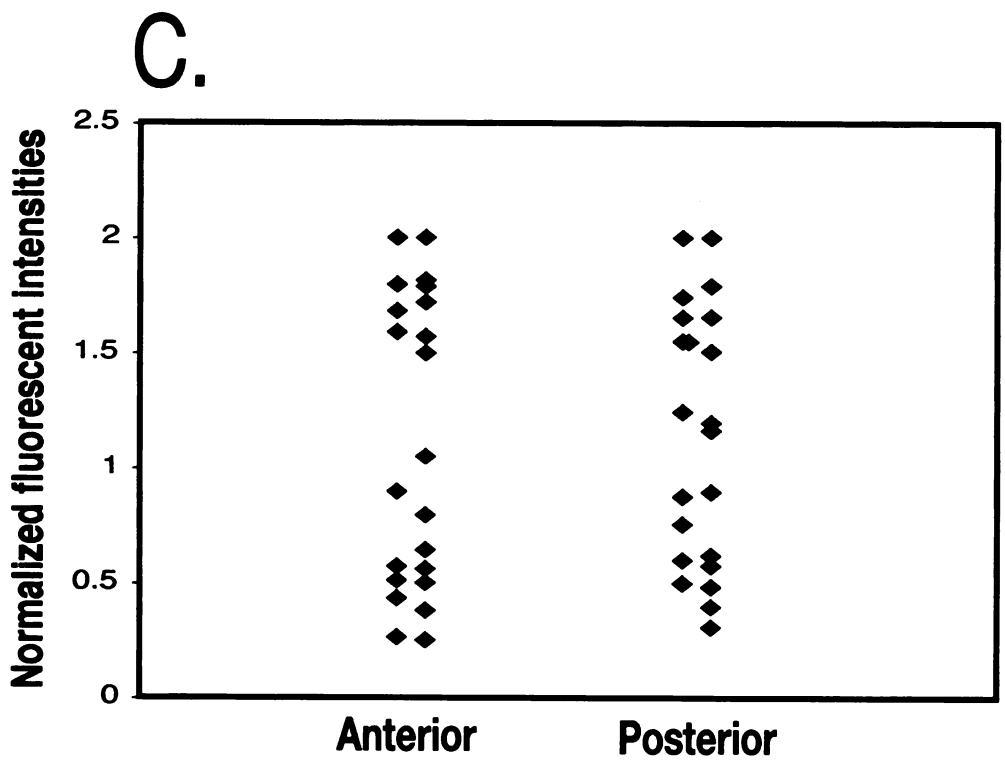


Figure II-3: Continued.



Methods for Data Collection:

In order to understand the functional relationship between transcriptional activator and repressor levels and mRNA output we need to create gene regulatory "maps" that describe the quantitative relationship between these elements.[20] We created a series of *lacZ* reporter genes regulated by the Giant transcriptional repressor and the Dorsal and Twist activators. These reporter genes are active in ventral regions of the embryo, and expression is interrupted in areas where Giant is expressed, depending on the arrangement of the binding sites (Fig. II-4A, B). We directly measure levels of the Giant repressor protein as an input value. The spatially and temporally varying levels of this protein provide in each embryo an entire set of values relevant to a gene regulatory map. Activator levels can be similarly measured, and in the case of Dorsal and Twist, these activators vary along the dorsal to ventral axis.[2] We focus on gene modules that test varying features of repressor binding sites, while holding Dorsal and Twist sites constant. Dynamic and spatially heterogeneous protein levels of transcriptional regulators make it imperative that we compare *lacZ* levels with corresponding levels of regulatory factors in nearby nuclei from which the mRNA originates.

Background:

Background intensity for the Giant channel is calculated by averaging the data from the middle (50-60% egg length) of the embryo along the ventral-dorsal axis. In cases where the curve shows linear behavior, the average of this middle stripe can be subtracted from the whole embryo, and if it shows parabolic behavior, then the background can be calculated and subtracted as previously described.[9] Background

intensity for the *lacZ* channel can be calculated similarly by using the dorsal parts of the embryo, where no *lacZ* is present.

Binning:

A pixel by pixel comparison between Giant, found in the nucleus, and the *lacZ* mRNA channel is not applicable because the mRNA accumulates in the spatially separated cytoplasm. In addition, we observe that *lacZ* mRNA accumulates in a punctuate pattern (a function of the specific reporter, as other mRNAs show a smoother, nonnuclear pattern, data not shown). This problem can be solved by measuring the approximate size of a "cell" in confocal images and covering the region of interest by boxes of the size of an average cell size. A 10x10 pixel box is the average cell size for our images.

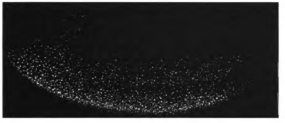
Relevant areas of data collection:

We then collect a series of data points representing the *lacZ* gene output in regions with similar levels of activators but varying levels of Giant protein. When not regulated by Giant, the *lacZ* expression pattern is almost constant from approximately 20-80% egg length (anterior-posterior) in the ventral part of the embryo (Fig. II-4A). When Giant is effective at repressing the *lacZ* reporter, gaps in this pattern are seen in anterior and posterior regions (Fig. II-4B). It has been reported that *twist* mRNA levels are not uniform from anterior to posterior[21] but this variation does not seem to be reflected in the output of our reporters, which are activated by Twist and Dorsal in a fairly uniform pattern in the mid-blastoderm stage. For this reason, in the regions that are exactly parallel to the ventral boundary, we assume that the combined activator effect is not changing for our synthetic enhancer constructs. We therefore measure correlated *lacZ*

Figure II-4: Sampling of integrated mRNA and Giant data

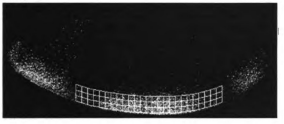
lacZ reporter constructs were introduced into *Drosophila* by germline transformation and mRNA detected by fluorescent imaging. Dorsal and Twist activators drive *lacZ* gene expression in ventral regions (A). When regulated by Giant, gaps in this pattern are evident (B). A sampling mesh is imposed upon the background subtracted image and values for Giant and *lacZ* are collected (C, D). The mesh size here is exaggerated for clarity. Sampling proceeds from anterior region of anterior stripe to posterior region of posterior stripe.

A.



B.

C.



D.

and Giant levels in "slices" that track along the ventral aspect of the embryo (Fig. II-4C, D). Curved slices, 10 pixels in width, are taken from ventral part of the embryo along anterior to posterior, following the boundary of the embryo. Each slice is divided into 10x10 pixel boxes along anterior to posterior to derive mean values for *lacZ* and Giant.

To accommodate images of embryos that are rotated around the anterior-posterior axis to different degrees, exposing more or less of the ventral surface where Dorsal/Twist are active, we vary the number of slices applied to each embryo so that our *lacZ* mRNA and Giant measurements reflect areas with nearly constant activator levels. To determine the number of slices used, we measure *lacZ* levels in central regions (50-60% egg length) that are not subject to Giant regulation and extend slices from dorsal to ventral until activator levels are determined to be limiting. This is accomplished by measuring *lacZ* expression in the 50-60% egg length portion of the embryo and averaging along the anterior-posterior axis, producing a plot that describes how *lacZ* is changing from ventral to dorsal. This data has a peak value at the ventral side of the embryo and decreases from ventral to dorsal. The slices start from the peak level of *lacZ* expression and end at 50% of the maximal *lacZ* expression.

Because the ventral *lacZ* expression is also affected by vector-mediated activation in anterior regions, as well as Torso-mediated regulation of Dorsal[22] in the poles, our slices start from a position of half of the maximal Giant intensity on the anterior side of the major anterior stripe, to half of the maximal Giant intensity of the posterior side of the posterior Giant stripe. The selected region still includes most of the region of Giant expression and removes portions of the data subject to extraneous influences (Fig. II-4C, D).

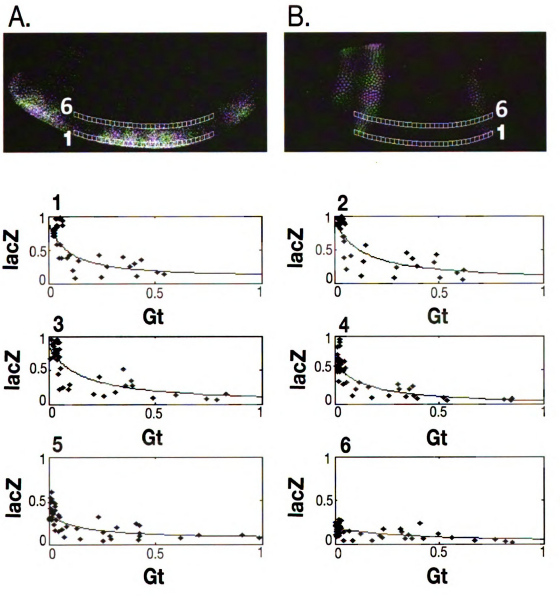
Normalization and regulatory plots:

The quantitation described above focuses on the *lacZ* levels as a function of the Giant concentration (Fig. II-5). The Giant data is normalized by dividing the 2-4h old embryo time interval studied here into eight segments as described in Jaeger *et al.* The average levels of the protein in each interval are used to normalize the Giant signals. In this way, comparisons of *lacZ* expression to relative Giant levels in a given embryo can be combined with data from other embryos into larger data sets. Because Dorsal and Twist activator proteins are relatively constant, each embryo can be normalized by dividing by its maximum intensity value. These normalized expression levels of Giant protein and *lacZ* mRNA are used to plot expression surfaces, showing the repressor signal and activator signal versus output (Fig.II- 5). For the visualization shown here, the data is presented as two dimensional plots, in which each box represents one slice from the ventral region of the embryo. As expected, with increasing Giant levels or decreasing activator levels, the *lacZ* signal is reduced. The resulting plot represents two dimensional slices of a gene regulatory surface that is unique to the specific regulatory region of interest.

Previous studies have demonstrated that the Giant protein can repress transgenes regulated by Dorsal and Twist, but the characterization of this relationship has never gone beyond a qualitative nature.[14] To characterize the quality of the data obtained from this system and assess its utility for quantitative modeling we measured the correlation between [Gt] and ln([*lacZ*]) using the Pearson correlation coefficient. For the 2gt.25.2T2D construct showing regulation by Giant, correlation coefficients for slice 1 to

Figure II-5: Gene regulatory representations from Giant regulated *lacZ* reporter gene

To generate gene regulatory plots, data derived from slices (*lacZ* mRNA channel, A and Giant protein channel, B) are plotted in a two dimensional space. Only two slices are shown for clarity. The plots numbered 1-6, show how robust *lacZ* levels present in ventral portions of the embryo are repressed in regions containing peak Giant levels such as in plot 1-4. In more dorsally located regions (plot 5-6) limiting levels of activators reduce *lacZ* expression regardless of Giant levels. The overlaid lines, obtained by fitting the data with a rational function show the trend of the relation between levels of Giant and *lacZ*.


A.
B.
6
1
6
1
1
2
3
4
5
6
lacZ
Gt
0
0.5
1

slice 6 are -0.74, -0.78, -0.82, -0.85, -0.65 and -0.67. The p-values for these correlation levels are all less than 0.0001, indicating statistical significance.

DISCUSSION

We describe a method to correlate the transcription factors and mRNA output for gene modeling purposes. We show that the levels of *lacZ* mRNA, and potentially the transcriptional repressor protein Giant, are proportional to fluorescent intensities, a critical basis for quantitative modeling. Our analysis also reveals that a suggested parabolic form of the background fluorescence in confocal images of early *Drosophila* embryos is evident most prominently in flattened specimens, with intact embryos exhibiting a more linear background. After appropriate background subtraction and normalization, these data are amenable to representation of gene regulatory surfaces that permit creation and validation of quantitative models of gene expression. In this way, we have constructed the foundations for modeling a *cis*-regulatory 'grammar' that applies to an important set of transcriptional regulators in the *Drosophila* blastoderm embryo. More generally, the image and data analysis techniques described in this paper can be generalized to understand the quantitative function of endogenous genes or gene networks in the Drosophila embryo or other well-characterized systems. Such an approach may also prove useful for engineered transcriptional regulatory elements employed for targeted expression of genes in therapeutic settings.

REFERENCES:

Jaeger J., Surkova S., Blagov M., Janssens H., Kosman D., Kozlov K.N., Manu, Myasnikova E., Vanario-Alonso C.E., Samsonova M., Sharp D.H., and Reinitz J. Dynamic control of positional information in the early *Drosophila* blastoderm. Nature **430**, 368, 2004.

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

Rivera-Pomar, R. and Jackle, H. From gradients to stripes in *Drosophila* embryogenesis: filling in the Gaps. Trends Genet **12**, 478, 1996.

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

Segal E., Raveh-Sadka T., Schroeder M., Unnerstall U., and Gaul U. Predicting expression patterns from regulatory sequence in *Drosophila* Segmentation. Nature **451**, 535, 2008.

Janssens H., Kosman D., Vanario-Alonso C. E., Jaeger J., Samsonova M., and Reinitz J. A high-throughput method for quantifying gene expression data from early *Drosophila* embryos. Dev Genes Evol **215**, 374, 2005.

Tomancak P., Beaton A., Weiszmann R., Kwan E., Shu S., Lewis S.E., Richards S., Ashburner M., Hartenstein V., Celniker S.E., and Rubin G.M. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. Genome Biol **3**, 88, 2002.

Luengo Hendriks C.L., Keranen S.V.E., Fowlkes C.C., Simirenko L., Weber G.H., DePace A.H., Henriquez C., Kaszuba D.W., Hamann B., Eisen M.B., Malik J., Sudar D., Biggin M.D., and Knowles D.W. Three-dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution 1: data acquisition pipeline. Genome Biol **7**, R123, 2006.

Myasnikova E., Samsonova M., Kosman D., and Reinitz J. Removal of background signal from in situ data on the expression of segmentation genes in *Drosophila*. Dev Genes Evol **215**, 320, 2005.

Small S., Blair A., and Levine M. Regulation of even-skipped stripe 2 in the *Drosophila* embryo. EMBO J **11**, 3147, 1992.

Kosman D., Mizutani C. M., Lemons D., Cox W. G., McGinnis W., and Bier E. Multiplex detection of RNA expression in *Drosophila* embryos. Science **305**, 846, 2004.

Kosman K., Small S. and Reinitz J. Rapid preparation of a panel of polyclonal antibodies to *Drosophila* segmentation proteins. Dev Genes Evol **208**, 290, 1998.

Szymanski P. and Levine M. Multiple modes of dorsal-bHLH transcriptional synergy in the *Drosophila* embryo. EMBO J **14**, 2229, 1995.

Hewitt G.F., Strunk B.S., Margulies C., Priputin T., Wang X.D., Amey R., Pabst B.A., Kosman D. Reinitz J. and Arnosti D.N. Transcriptional repression by the *Drosophila* giant protein: *cis* element positioning provides an alternative means of interpreting an effector gradient. Development **126**, 1201, 1999

Huggett J., Dhehada K., Bustin S., and Zumla A. Real-time RT-PCR normalization; strategies and considerations. Genes Immun **6**, 279, 2005.

Vandesompele J., De Preter, K., Pattyn F., Poppe B., Van Roy N., De Paepe A., and Speleman F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol **3**, 34, 2002.

Eldon E., and Pirrotta V. Interactions of the Drosophila gap gene giant with maternal and zygotic pattern-forming genes. Development **111**, 367, 1991.

Myasnikova E., Samsonova A., Kozlov K., Samsonova M., and Reinitz, J. Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods. Bioinformatics **17**, 3, 2001.

Damle S., Hanser B., Davidson E. H., and Fraser S.E. Confocal quantification of *cis*-regulatory reporter gene expression in living sea urchin. Dev Biol **299**, 543, 2006.

Setty Y., Mayo A.E., Surette M.G., and Alon U. Detailed map of *cis*-regulatory input function. PNAS **100**, 7702, 2003.

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

Rusch J., and Levine M. Regulation of the dorsal morphogen by the Toll and torso signaling pathways: a receptor tyrosine kinase selectively masks transcriptional repression. Genes Dev **8**, 1247, 1994.

Chapter III

Deciphering a transcriptional grammar: modeling short-range repression in the *Drosophila* embryo

Walid D. Fakhouri, Ahmet Ay, Rupinder Sayal, Jacqueline Dresch, Evan Dayringer, Chichia Chiu and David N. Arnosti submitted to Molecular Systems Biology

ABSTRACT

Systems biology seeks a genomic-level interpretation of transcriptional regulatory information represented by patterns of protein binding sites. Obtaining this information without direct experimentation is challenging; minor alterations in binding sites can have profound effects on gene expression, and underlie key aspects of disease and evolution. Quantitative modeling offers an alternative path to develop a global understanding of the transcriptional regulatory code. Recent studies have focused on endogenous regulatory sequences; however distinct enhancers differ in many features, making it difficult to generalize to other *cis*-regulatory elements. We applied a systematic approach to simpler elements, and present here the first quantitative analysis of short-range transcriptional repressors, which play central roles in metazoan development. Our fractional occupancy-based modeling uncovered unexpected features of these proteins' activity that allow accurate predictions of regulation by the Giant, Knirps, Krüppel and Snail repressors, including modeling of an endogenous enhancer. This study provides essential elements of a transcriptional grammar that will allow extensive analysis of genomic information in *Drosophila melanogaster* and related organisms.

INTRODUCTION

The rapid increase in sequenced genomes has provided an extensive "parts lists" of organisms, however deeper understanding of genomic information that includes gene regulatory functions is critical to understanding the dynamic activity of biological systems. Subtle changes in regulatory elements are often involved in hereditary diseases, population differences and the evolution of morphological novelties (Carroll *et al*, 2001). Comparative studies have demonstrated that regulatory regions can retain function over

large evolutionary distances, even though the DNA sequences are divergent and poorly alignable (Ludwig & Kreitman, 1995; Hare *et al*, 2008). The flexibility in arrangement of binding sites is not unlimited, however. For instance, the effectiveness of short-range transcriptional repressors that play key roles in *Drosophila* development is strongly influenced by activator-repressor distances (Gray *et al*, 1994; Arnosti *et al*, 1996a; Kulkarni & Arnosti, 2005).

The *Drosophila* blastoderm embryo provides an ideal setting for the analysis of transcriptional enhancers; the cascade of maternally and zygotically supplied transcription factors has been extensively investigated at a molecular level, and many DNA regulatory elements have been identified and functionally dissected. In this system, genes with complex expression patterns are controlled by multiple enhancers, whose modular function depends on the local action of repressor proteins (Small *et al*, 1993). The blastoderm embryo has been used for quantitative analysis of gene expression by reaction diffusion, Boolean, and fractional occupancy modeling (Jaeger et al, 2006; Sánchez & Thieffry, 2001; Segal *et al*, 2008). Fractional occupancy models draw from simple biophysical principles and statistical physics to predict the overall readout of endogenous enhancers (Bintu *et al*, 2005a, 2005b). In these models, parameters include the binding affinity of transcription factors to the DNA and cooperativity between proteins. Such models assume that gene regulation is dictated largely by the equilibrium binding of transcription factors to the DNA, without explicitly modeling events such as chromatin modifications and RNA polymerase phosphorylation.

Simple prokaryotic systems provide a tractable setting for quantitative studies, and fractional occupancy models have been applied to the *lac* operon in *E. coli* and the

lysis/lysogeny switch of phage lambda (Von Hippel *et al*, 1974; Ackers *et al*, 1982; Shea & Ackers 1985, Vilar & Leibler, 2003). Use of these models in eukaryotes is more problematic, given the higher degree of enhancer complexity in eukaryotic systems, but *Drosophila* enhancers have been treated by fractional occupancy models that account for factor spacing and recruitment of co-regulators (Janssens *et al*, 2006; Zinzen *et al*, 2006). These models can reproduce the behavior of specific enhancers, but a major limitation of fractional occupancy modeling of endogenous enhancers is that models of a single regulatory region, may not generally apply to other elements. In studies of multiple enhancers, the parameter estimation has been difficult, as the different architecture of distinct enhancers, even those regulated by the same proteins, makes it difficult to know which parameters (number of bindings sites, relative arrangements etc.) are key to determining the particular activity of an enhancer (Segal *et al*, 2008). As we describe here, a more systematic approach is necessary to parse the contributions of individual physical features to enhancer activity.

One particular area that has been inadequately explored is the key role played by repressor proteins. Giant, Knirps, and Krüppel are regionally deployed short-range repressor proteins that bind to and control the patterning of pair-rule genes such as *even-skipped*. Previous studies showed that precise positioning of short-range repressors on an enhancer can be used to generate the appropriate expression pattern in a morphogenetic field where the concentration of these repressors are used to set gene expression thresholds (Hewitt *et al*, 1999; Clyde *et al*, 2003). Thus, the flexibility of enhancer architecture incorporating these proteins is constrained by some distance limitations. Our previous study demonstrated that activator-repressor stoichiometry and arrangement of

binding sites also influence the overall readout of developmental enhancers (Kulkarni & Arnosti, 2005). To build tools able to accurately predict the function of novel enhancer sequences, we recognized a need to quantitatively measure the specific contributions of these factors to overall enhancer function. Here we describe the creation and quantitative assessment of a well-defined set of transcriptional regulatory modules in the *Drosophila* embryo, in which individual aspects relating to repressor-activator spacing, stoichiometry and arrangement are systematically explored. Using quantitative data from these genes, we apply a fractional occupancy approach to model the interaction of short-range repressors with endogenous transcriptional activators. We show that this approach can correctly decipher the transcriptional regulatory code of endogenous enhancers, pointing the way to a general approach for unlocking the transcriptional regulatory information of genomes.

MATERIALS AND METHODS

Reporter genes

The binding motifs for the Giant, Krüppel and Knirps short-range repressors and the Twist and Dorsal activators used in this study were characterized in previous studies (Szymanski & Levine, 1995; Hewitt *et al*, 1999; Kulkarni & Arnosti, 2005).

Regulatory modules were constructed in pBluescript KS(+) using the EcoRI, BamHI, XbaI and SacII restriction sites, amplified by PCR using T3 and T7 primers, and amplicons were digested with EcoRI and SacII and subcloned into the compatible sites of C4PLZ (Wharton & Crews, 1993). Gene 1 contains two Giant binding sites inserted

between EcoRI and BamHI, two Twist sites inserted between BamHI and XbaI, and two Dorsal binding sites between XbaI and SacII.

Gene 2 includes a 25bp spacer inserted between Giant and Twist sites using BamHI. For genes 4, 6, 7 and 8, the same 25bp spacer was concatemerized and inserted at BamHI. For genes 3 and 5, a 35bp or 60bp spacer was inserted at BamHI, between the Giant and Twist binding sites. Gene 9 contains a single Giant binding site inserted between EcoRI and BamHI. Gene 10 was constructed by digestion of the parent gene 1 pBluescript plasmid with EcoRI and insertion of the single Giant binding site, preserving a single 5' EcoRI site. For genes 12, 13, 14, 15 and 19 the same strategy was used to insert an extra Giant binding site 3' of the Dorsal sites using SacII. For genes 13 and 15, in which the binding sites are moved away from the basal promoter of *lacZ* reporter gene, a 340bp spacer was amplified from the coding region of *knirps* gene and inserted into the SpeI of C4PLZ plasmid (Kulkarni & Arnosti, 2005). A weaker Giant site was tested in gene 20. The sequences for all oligos used are shown in Appendix A Table III. All gene cassettes were confirmed by sequencing, and at least 5 transgenic lines of each gene were analyzed by *in situ* hybridization for *lacZ* expression pattern. Lines showing enhancer trapping were not included in the analysis. Fixed embryos from two to three transgenic lines of each gene were used for confocal laser scanning microscopy (CLSM).

Image processing

A five step procedure was applied to all embryo images as described in Ay *et al* (2008), involving binary image generation, rotation, outlier removal, background subtraction and normalization. Non-specific signals were not observed in Krüppel and

Knirps channels, so no outlier removal was required for these channels. Background intensity for the Giant channel was calculated by fitting a parabola along the dorsal-ventral axis to the average of the data from the middle (50–60% egg length) of the embryo along the anterior-posterior axis, a region lacking any Giant expression. Background intensity for the Krüppel and Knirps channels was calculated and removed similar to Giant channel, by using the data from the middle (75–85% and 25–35% egg length respectively) of the embryo along the ventral–dorsal axis.

The Giant channel was normalized so that embryos in the same age group with similar rotations have the same total average signal in anterior and posterior stripes. The *lacZ* channel was normalized using the average signal in a region bounded by the 50-60% egg length (anterior-posterior axis) and the peak to 60% of the peak *lacZ* signal (dorsal-ventral axis). Krüppel and Knirps channels were normalized similarly to the Giant channel. We also tested normalization scheme for Giant levels that used the mean of the peak 0.25% values set to 153 of 255 by linear transformation (The value 153 was chosen empirically to minimize the number of saturated pixels). The former scheme was more effective in normalizing Giant levels, but the [*lacZ*] vs. [Gt] plots were not significantly altered, possibly due to the higher level of noise in the *lacZ* channel.

Data set

A total of 769 embryos bearing *lacZ* reporter gene constructs regulated by the Giant repressor protein were analyzed and an additional 45 and 88 were analyzed for genes regulated by Krüppel and Knirps respectively. Genes 6, 7, 8 and 18 were not used in the quantitative modeling because no Giant repression was ever observed, and an

ectopic modulation of the reporter gene by unknown factors in a fraction of the imaged embryos made this data especially noisy. All primary data are available on our server at: www.bch.msu.edu/faculty/arnosti.htm

Parameter estimation

There are two main groups of approaches for parameter fitting; local and global parameter estimation techniques. Both groups might get stuck at local optima, but the chance of finding the global optimum is greater with global parameter estimation techniques. We compared three global techniques, namely evolutionary strategy (Runarsson & Yao, 2005), genetic algorithm (The MathWorks, Inc.) and simulated annealing (The MathWorks, Inc.) .These methods have been shown to be suitable for parameter estimation in biological systems in several studies (Mendes & Kell, 1998; Moles *et al*, 2003; Fomekong-Nanfack *et al*, 2007). We tested these estimation techniques by running them on a synthetic data set produced by calculating our model output for assigned parameters; such parameter estimation is a common technique for validation (Appendix A Figures 1-4) (Moles *et al*, 2003). We further checked their robustness on the synthetic data by introducing noise into the data. Strikingly different results were obtained when these approaches were tested on the model. After 100 simulated runs, the genetic algorithm and simulated annealing techniques were unable to accurately predict the correct parameters, even with no noise introduced into the data set (Appendix A Figures 2-3). In contrast, the evolutionary strategy algorithm was able to reproducibly predict the correct parameter set (Appendix A Figures 1 and 4). To determine the robustness of the parameter estimation techniques to noise, the estimation

algorithms were also run on the synthetic data with noise of normal distribution. Not surprisingly, the variance for genetic algorithm and simulated annealing algorithm was high at all levels of noise. Even at a noise level of 30%, the variation for the evolutionary strategy approach was lower than the other two methods (Appendix A Figure 4). It is possible that genetic algorithm or simulated annealing may be capable of producing better results if implemented differently, however if the choice of parameter estimation is not examined carefully, a broad range of values may be produced, leading to the conclusion that there are no optimal values.

Schemes

Distinct forms of our model were implemented in which the quenching parameters were grouped in different 'bins' of distances. For distances more than 81bp quenching efficiencies of the repressors are taken as 0, motivated by our genes 6-8, which shows no repression.

Schemes 1, 2 and 8: q_1(6bp), $q_2(28-41\text{bp})$, $q_3(50-56\text{bp})$, $q_4(63-66\text{bp})$, q_5(78bp), q_6(6bp from 3' end of activators).

Schemes 3, 4 and 9: q_1(6bp), $q_2(28-41\text{bp})$, $q_3(50-53\text{bp})$, $q_4(56-66\text{bp})$, q_5(78bp), q_6(6bp from 3' end of activators).

Schemes 5: q_1(6bp), $q_2(28-31\text{bp})$, $q_3(41-50\text{bp})$, $q_4(53-56\text{bp})$, $q_5(63-66\text{bp})$, q_6(78bp), q_7(6bp from 3' end of activators).

Schemes 6: q_1(6bp), q_2(28 – 31bp), q_3(41bp), q_4(50 – 56bp), q_5(63 – 66bp), q_6(78bp), q_7(6bp from 3' end of activators).

Schemes 7: q_1(6bp), q_2(28bp), q_3(31bp), q_4(41bp), q_5(50bp), q_6(53bp), q_7(56bp), q_8(63bp), q_9(66bp), q_{10}(78bp), q_{11}(6bp from 3' end of activators).

We also tried different expressions of cooperativity. In schemes 1, 3, 5, 6 and 7 only one parameter is used for Giant-Giant cooperativity. In schemes 2, 4, 8 and 9 two parameters are used for Giant-Giant cooperativity, where the first parameter describes cooperativity of Giant proteins with 10bp distance and the second describes cooperativity for 32bp distance. In schemes 2 and 4 we described the cooperativity of observing all three Giant proteins on the DNA as the summation of the two cooperativity parameters, and in schemes 8 and 9 as the multiplication of the two cooperativity parameters.

Derivation of the model for gene 1

We express efficiency of the activator group bound and not bound respectively as E_A and E_N, and efficiency of the Giant repressor as E_{Gt}. We represent the efficiency vector that represents efficiency for each state of activator set and Giant repressors as E, the state vector of activator set and Giant repressors as F, the regulatory function that transforms each efficiency vector input to transcription level as T, and total steady state transcription level as Ex. The probability of each state of those proteins on the DNA can be calculated. Because the activator binding sites do not vary within the genes tested here, the Dorsal and Twist activators are not parameterized, and are considered as one group. We set $S_A[A]$ equal to 100 with the assumption that in the

absence of Giant repressor protein, the activators are fully functional. A set of eight equations describes all possible states of this gene. For simplification in the following formulas,

$$Z = (1 + S_A[A])(1 + 2S_R[Gt] + C(S_R[Gt])^2).$$

No activator and repressor bound: $F_N = \frac{1}{Z}$

Activator set is bound: $F_A = \frac{S_A[A]}{Z}$

Proximal Giant to the activator set is bound: $F_{Gt_1} = \frac{S_R[Gt]}{Z}$

Distal Giant to the activator set is bound: $F_{Gt_2} = \frac{S_R[Gt]}{Z}$

Activator set and proximal Giant to the activator set is bound: $F_{AGt_1} = \frac{S_A[A]S_R[Gt]}{Z}$

Activator set and distal Giant to the activator set is bound: $F_{AGt_2} = \frac{S_A[A]S_R[Gt]}{Z}$

Both Giant repressors are bound: $F_{Gt_1Gt_2} = \frac{C(S_R[Gt])^2}{Z}$

Activator set and both Giant repressors are bound: $F_{AGt_1Gt_2} = \frac{S_A[A]C(S_R[Gt])^2}{Z}$

Then the states vector of one activator set and two repressors can be written as:

$$F = [F_N, F_A, F_{Gt_1}, F_{Gt_2}, F_{AGt_1}, F_{AGt_2}, F_{Gt_1Gt_2}, F_{AGt_1Gt_2}]$$

In the expressions above we stated all the Boltzmann states of an enhancer with one activator set and two repressor binding sites. We claim that the binding of the repressors modulates the probability of states with activators and repressors simultaneously bound by a quenching factor. For example binding of the activator A and repressor R_1 simultaneously is reduced by a factor of $(1-q_1)$. We can modify the probability of states, in the following way:

$$\tilde{F} = [F_N, F_A, F_{Gt_1}, F_{Gt_2}, F_{AGt_1}(1-q_1), F_{AGt_2}(1-q_2), F_{Gt_1Gt_2}, F_{AGt_1Gt_2}(1-q_1)(1-q_2)]$$

Next we calculate the total efficiency of the enhancer when factors are bound on the DNA. We model cooperativity between Giant sites, which might be for example due to cooperative cofactor recruitment, with an additive function, so the total efficiency of Giant repressors bound to the DNA at the same time is $E_{Gt_1Gt_2} = w_1^{Gt} E_{Gt_1} + w_2^{Gt} E_{Gt_2}$ where w_1^{Gt} and w_2^{Gt} are the cooperativity terms after binding. The efficiency of one activator set and two repressors are expressed in the following way; each term representing one state of the all possible states:

$$E = [E_N, E_A, E_{Gt_1}, E_{Gt_2}, E_{AGt_1}, E_{AGt_2}, E_{Gt_1Gt_2}, E_{AGt_1Gt_2}]$$
$$= [E_N, E_A, E_{Gt_1}, E_{Gt_2}, E_A + E_{Gt_1}, E_A + E_{Gt_2}, w_1^{Gt} E_{Gt_1} + w_2^{Gt} E_{Gt_2},$$
$$E_A + w_1^{Gt} E_{Gt_1} + w_2^{Gt} E_{Gt_2}]$$

Expression contributions from each state are added to obtain the total expression:

$$Ex = \sum_i \tilde{F}_i T(E_i)$$

If we set the following simple assumptions $E_N = 0$, $E_A = 10$, $E_{Gt_1} = 0$, $E_{Gt_2} = 0$ and $T(x) = \frac{1}{1 + e^{5-x}}$ the total expression of the enhancer with 1 activator set and 2 repressor binding sites can be written as:

$$Ex \approx \frac{S_A[A]}{1 + S_A[A]} \times \frac{1 + (2 - q_1 - q_2) S_R[Gt] + C(1 - q_1)(1 - q_2)(S_R[Gt])^2}{1 + 2S_R[Gt] + C(S_R[Gt])^2}$$

Expression functions (Ex) for all cases are shown in Appendix A Table I. Further details about the model are explained in the supplementary material (Appendix A).

Modeling endogenous enhancer sequences

We made the following assumptions to simplify the parameter estimation for modeling of the *rho* NEE: (1) we model activity of the NEE in the mesoderm, where Dorsal and Twist levels are high, and Snail is present at uniform levels. We used values for expression contribution of Dorsal and Twist as +5 each, and for Snail, -5. We also

carried out parameter estimation with values of +3 or +7 for activators and -3 or -7 for Snail, and obtained essentially equivalent results. (2) We set quenching parameters to those obtained from our modeling, as shown in Figure III-4, on the reasonable assumption that these are functionally equivalent among short-range repressors, (3) To reduce the number of possible parameters, we only included cooperative interactions between factors that are nearest neighbors, and are located within 25 bp of each other. (4) We allow that the relative effectiveness of repression with four Snail sites might be higher than that seen with one or two, and stipulate ranges of repression in which parameter space is investigated (Figure III-8). (5) We set ranges for cooperativity and scaling factors from 1-100. (6) For each transcription factor, we took the score of the strongest site among all those that bind that transcription factor as a free parameter and constrain the other values by treating the PWM score as a free energy of binding (Stormo, 2000). We used PWMs created from FlyReg database by Daniel A. Pollard, which are available at: htt://www.flyreg.org/. As an example, the two Twist sites differ considerably in terms of their match to a consensus PWM, with Twist site #2 predicted to have a forty seven-fold lower score than Twist site #1, although it still has a considerably higher score than background sequences.

RESULTS

Gene modules

We set out to map regulatory surfaces of genes controlled by short-range repressors; these surfaces show the functional relationship of activator/repressor input and gene expression output (Figure III-1). Such regulatory surfaces reflect evolutionary

forces that shape gene output, as demonstrated for the *lac* operon (Setty *et al*, 2003; Mayo *et al*, 2006). The design of the enhancers responding to short-range repressors accommodates sensitive distance and binding site parameters within a flexible design framework (Clyde *et al*, 2003; Kulkarni & Arnosti, 2005).

The output of a model of a particular configuration of transcription factor binding sites should lead to a regulatory surface that allows mapping of known values of regulatory factors, such as Dorsal and Twist activator protein levels, and Giant repressor protein levels, through this surface to produce an expected regulatory outcome (Figure III-1).

To carry out this scheme on a practical level, we created a series of genes to test in a systematic fashion the effect of parameters affecting repression. The quantitative measurement of these genes was used to create a database suitable for quantitative modeling, identification of parameters related to repressor activity, and analysis of endogenous regulatory elements (Figure III-1E). We used endogenous activators and repressors that are active in the blastoderm embryo. A convenient juxtaposition of anterior-posteriorly expressed repressor proteins Giant, Krüppel or Knirps are superimposed on the patterns derived from activators working on the dorsal-ventral axis to generate readouts as shown in Figure III-1. This design permits the simultaneous monitoring of repressed and unrepressed states in a single embryo. Twenty-seven *P*-element based genes d into the *Drosophila* germline to produce stably integrated *lacZ* reporters. We tested multiple lines for each; position effects had some effect on overall expression levels, but not on relative repression effectiveness. As described below, activator signals are normalized before parameter estimation and modeling, removing this

Figure III-1: Transformation of DNA sequence and protein information by gene modeling. (A) An enhancer with three repressors (red squares) and four activators (green circles) is modeled, to generate the gene expression surface shown in (B). The axes represent normalized activator, repressor and gene activity levels. (C) A *Drosophila* embryo with Giant repressor (red stripes) and Dorsal activator (green) staining is shown. Each embryo provides a diversity of potential inputs to the regulatory element: the white arrow points to a region where activator levels are high and repressor levels are low. The black arrow points to a region where both activator and repressor levels are low. The white triangle points to a region where activator and repressor levels are both high, and the black triangle points to a region where repressor levels are high and activator levels are low. (D) Output of regulatory element shown in (A), which mirrors values from (C) being mapped through surface shown in (B). (E) Formal scheme of data collection, analysis and modeling.

A

B

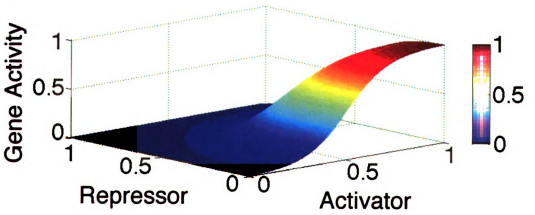
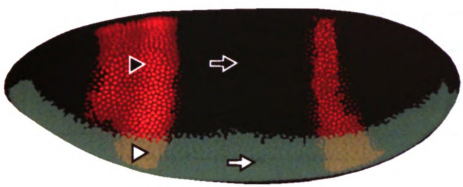


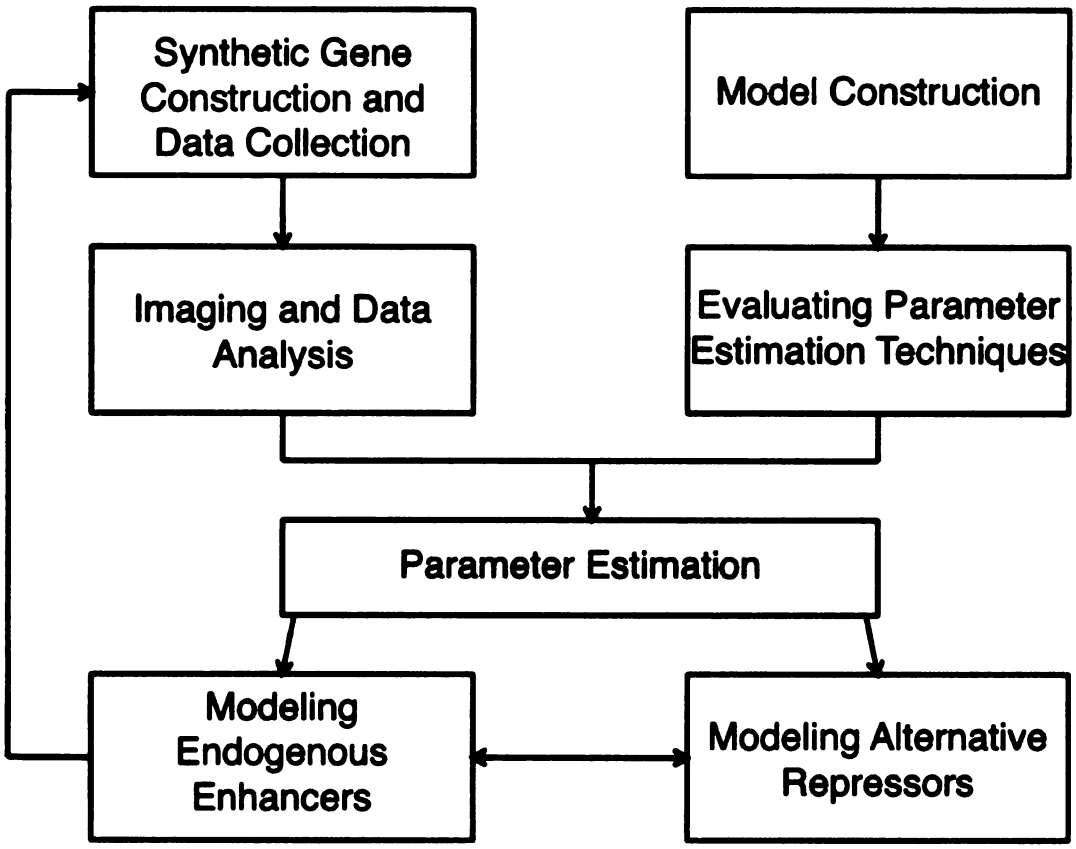
Figure III-1: Continued.

C



D

Figure III-1: Continued.



potential source of variability. Based on previous studies, we knew that spacing between activators and repressors would be a critical element to model, thus a series of genes (1-8, Figure III-2) tested variable distances between Giant repressor binding sites and the nearest Twist activator sites. As revealed by conventional *in situ* staining, repression effectiveness was markedly attenuated by this increase in spacing. Genes for which the most proximal binding site for Giant was located at least 81 bp from the nearest Twist site failed to show any repression (genes 6-8, Figure III-2). A gene containing a single Giant binding site adjacent to the Twist activators was weakly repressed, consistent with earlier reports (Hewitt *et al*, 1999), and this repression was also found to be distance-dependent (genes 9, 16).

Increasing the number of binding sites to three (genes 10, 17, 18) appeared to generate an especially effective repression context, one that was similarly susceptible to distance effects; at this level of resolution, it was not clear whether the distance function is appreciably different with different numbers of repressors. We also tested the effect of arranging the repressors in a distinct pattern so that some sites were located 3' of the activator cluster, adjacent to the Dorsal activator sites. In this way, we were able to test whether overall stoichiometry of repressors to activators was the sole determinant of repression effectiveness when binding sites are close to the activators. We noted that different distributions of two or three sites appeared to yield similar results, whether all sites were located 5' of the activator cluster, proximal to the Twist activator sites, or with some of the Giant repressor sites located 3' of the activator cluster, adjacent to Dorsal (genes 12, 14, 19). Insertion of a 340 bp neutral spacer sequence between the transcription factor cluster and the basal promoter did not change the pattern of gene

Figure III-2: Structures of genes assayed to determine context dependence of short-range repressor activity, and representative *in situ* images showing *lacZ* activity. Mid blastoderm embryos are oriented dorsal up, anterior to the left.

1. 2Gt.2Tw.2Dl

2. 2Gt.25.2Tw.2Dl

3. 2Gt.35.2Tw.2Dl

4. 2Gt.50.2Tw.2Dl

5. 2Gt.60.2Tw.2Dl

6. 2Gt.75.2Tw.2Dl

7. 2Gt.100.2Tw.2Dl

8. 2Gt.125.2Tw.2Dl

9. 1Gt.2Tw.2Dl

10. 3Gt.2Tw.2Dl

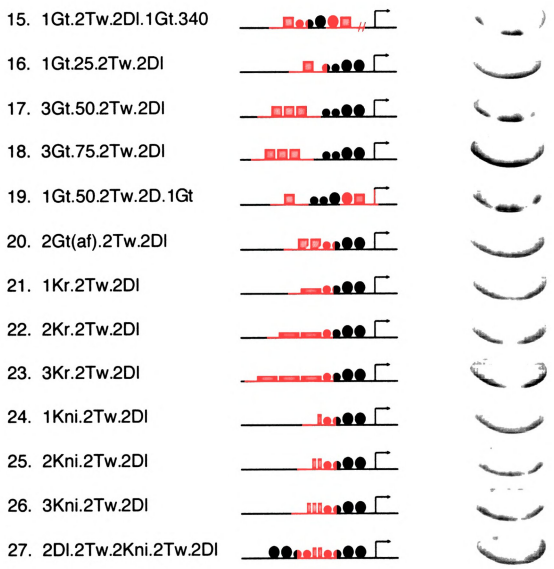
11. 2Dl.2Tw.2Gt.2Tw.2Dl

12. 2Gt.2Tw.2Dl.1Gt

13. 2Gt.2Tw.2Dl.1Gt.340

14. 1Gt.2Tw.2Dl.1Gt

Figure III-2: Continued.



expression, suggesting that the repressor is not acting directly on the basal promoter in this context (genes 12 vs. 13; 14 vs. 15). Most blastoderm enhancers characterized for these regulatory proteins are located some distance from transcriptional start sites, thus the distance independence of these modules mimics the activity of endogenous enhancers. We furthermore tested the effect of increasing the number of activators located in the vicinity of the repressors (genes 11, 27) and found that repression effectiveness was little compromised in the case of Giant, but appeared to be attenuated in the case of the weaker Knirps repression. Weaker binding sites for Giant produced attenuated repression, as expected (gene 20). Finally, a series of genes with increasing numbers of binding sites for Knirps and Krüppel allowed for direct comparison of repressor effectiveness and effects of stoichiometry (genes 21-26); as noted for Giant, more sites were generally more effective, but overall repression effectiveness of Knirps was lower. This difference may be attributed to weaker binding sites, lower absolute levels of the protein, or protein activity, as discussed below. The quantitative analysis of these genes was followed by quantitative measurements, described below.

Image Processing and Data Analysis

To simplify modeling, we initially restricted our measurements to the regions of the embryos containing peak levels of the Dorsal and Twist activators, which were identified as ventral regions expressing >60% of peak *lacZ* levels. To identify gene responses to varying repressor levels, we generated correlated Giant protein/*lacZ* mRNA plots (Figure III-3). This step involved a series of image processing procedures, as described in Ay *et al* (2008). We first identified and subtracted non specific signals

("outliers") observed in the Giant channel, then identified and subtracted background from each embryo. Background intensities for the *lacZ* and Giant channels were subtracted using average values from regions lacking activators and repressors. Next, we normalized the Giant channel for similarly aged and oriented embryos. The *lacZ* channel was normalized using the average signal in a region defined by 50-60% egg length of the embryo (anterior-posterior) and peak to 60% of the peak (dorsal-ventral).

Our data set comprises expression data from 20 *lacZ* reporter genes regulated by Giant, 3 *lacZ* reporter genes regulated by Krüppel and 4 *lacZ* reporter gene constructs regulated by Knirps. Over 900 blastoderm embryos were quantified to aid in parameterization of repressor and *lacZ* expression. Images were processed as described above and [*lacZ*] vs. [repressor] (Giant, Krüppel or Knirps) plots were created. The relative levels of gene expression as a function of repressor protein were plotted for individual images and compiled into composite plots (Figure III-3) (Ay *et al*, 2008). These plots were used to infer *cis*-regulatory rules by fractional occupancy models as described below.

Fractional Occupancy Modeling

Fractional occupancy models of transcriptional regulatory regions enumerate all possible 'states' of an enhancer based on potential transcription factor-DNA interactions, and then calculate the probability of a gene firing as the fraction of the 'successful' states, i.e. those with activators bound, and without excessive interference by repressors (Bintu *et al*, 2005a; Janssens et al, 2006; Zinzen *et al*, 2006; Segal *et al*, 2008). To capture the key role of short-range repressors on activator elements, we used a modified fractional

site occupancy model that explicitly accounts for distances between activators and short-range repressors, as well as cooperativity and binding affinity of short-range repressors. We allow for change in repression with distance but make no *a priori* assumptions about how the repression efficiency changes.

For a general description of our model, we employ three parameter types: S_R, a repressor scaling factor, indicating the potency of the repressor, C, representing cooperativity between repressor proteins binding to sites that are close together, and q, representing the distance-dependent "quenching" efficiency of the short-range repressors. In genes assayed here, the activator binding sites do not vary; therefore additional parameters representing activator potency or binding cooperativity are not required. A more sophisticated general model incorporating these features is described below for endogenous sequences.

To apply this model to one of our genes, 2Gt.2Tw.2Dl (gene 1), we express normalized activator and Giant repressor concentrations respectively as $[A]$ and $[Gt]$, activator and Giant repressor scaling factors as S_A and S_R (which represent binding affinity and concentration scaling combined into one scaling factor) $(1 \le S_A, S_R \le 100)$, and cooperativity between Giant repressor proteins for binding to DNA as C $(0.1 \le C \le 100)$. Quenching, the distance-dependent repression efficiency, is represented by q_1 and q_2 for the two Giant repressors in this gene $(0 \le q_1, q_2 \le 1)$. As derived in Materials and Methods, the expression of this gene when fully bound by activators and repressors will be:

Figure III-3: Representative [*lacZ*] vs. [Gt] plots.

(A) Structures of three genes assayed (1, 9 and 10). (B-C) Representative embryos imaged for Giant protein and *lacZ* reporter gene activity. (D) The data from multiple confocal embryo images was processed and compiled to provide normalized reporter gene [*lacZ*] vs. normalized repressor [Gt].

A
B
C
D
2.5
[lacZ]
0
[Gt]
1
2.5
[lacZ]
0
[Gt]
1
2.5
[lacZ]
0
[Gt]
1

$$Ex \approx \frac{S_A[A]}{1+S_A[A]} \times \frac{1+(2-q_1-q_2)S_R[Gt]+C(1-q_1)(1-q_2)(S_R[Gt])^2}{1+2S_R[Gt]+C(S_R[Gt])^2}$$

Comparable expressions are generated for each of the genes (Appendix A Table I).

Parameter Estimation

Parameter estimation is a critical step in implementation of modeling. Here we infer relationships of regulatory factors from the transcriptional output of reporter genes, which involves solving an inverse problem that may not have a unique solution. For this reason, no parameter estimation technique works well for all problems, so the choice of the parameter estimation technique is critical. In the transcriptional modeling literature, however, this facet has not been explicitly treated by modelers, and the choice of the parameter estimation technique is often not validated (Janssens *et al*, 2006; Zinzen *et al*, 2006; Segal *et al,* 2008). We tested three popular global parameter estimation techniques, evolutionary strategy (Runarsson & Yao, 2005), genetic algorithm (The MathWorks, Inc.) and simulated annealing (The MathWorks, Inc.) and found that for our model the evolutionary strategy produced superior results as described in Appendix A Figures 1-4.

Testing/Implementing Nine Forms of the Model

To analyze the quantitative data obtained from the embryos, we built nine forms of the model featuring increasing complexity in terms of number of parameters used; the models differ in their treatment of cooperativity and quenching distance. In the simpler case, a single parameter represents cooperativity between adjacent Giant repressor binding sites, as well as the interaction of all three sites involved in genes 10, 17 and 18.

Alternatively, we also employed a more complex treatment in which adjacent sites are fit to C_1 and sites separated by intervening Giant sites are fit to C_2. Similarly, quenching efficiency parameters of repressors can be defined either as unique parameters for each distance or as parameters for a range of distances, as described in Materials and Methods.

We show a pictorial description of the parameter assignments for scheme 2, a simpler form, in Table III-1. Appendix A Table II provides a pictorial description of the parameter assignments for all schemes.

We compared the nine schemes as explained in model validation section below. As judged by the error comparison, schemes 1-4, 8 and 9 work better than schemes 5-7 in this data set, probably due to the smaller number of parameters (Appendix A Figure 6). Here for further analysis we showed the results of scheme 2. The results of the schemes 1, 3-6, 8 and 9 were comparable, suggesting that conclusions drawn from scheme 2 are representative (Appendix A Figure 5).

Model Predictions

Previously identified qualitative relationships about quenching and cooperativity/activity provide the backdrop for this work; the quantitative relationships presented here constitute the heart of this study, obtained after modeling our quantitative data set. It was striking that certain qualitative and quantitative insights became apparent only after analysis of the complete data set; these were not relationships that would necessarily be evident by inspection of individually stained embryos in Figure III-2. First, our model predicts rather modest levels of Giant-Giant cooperativity, greater than simply additive but lower than previous estimates (Figure III-4A) (Segal *et al*, 2008).

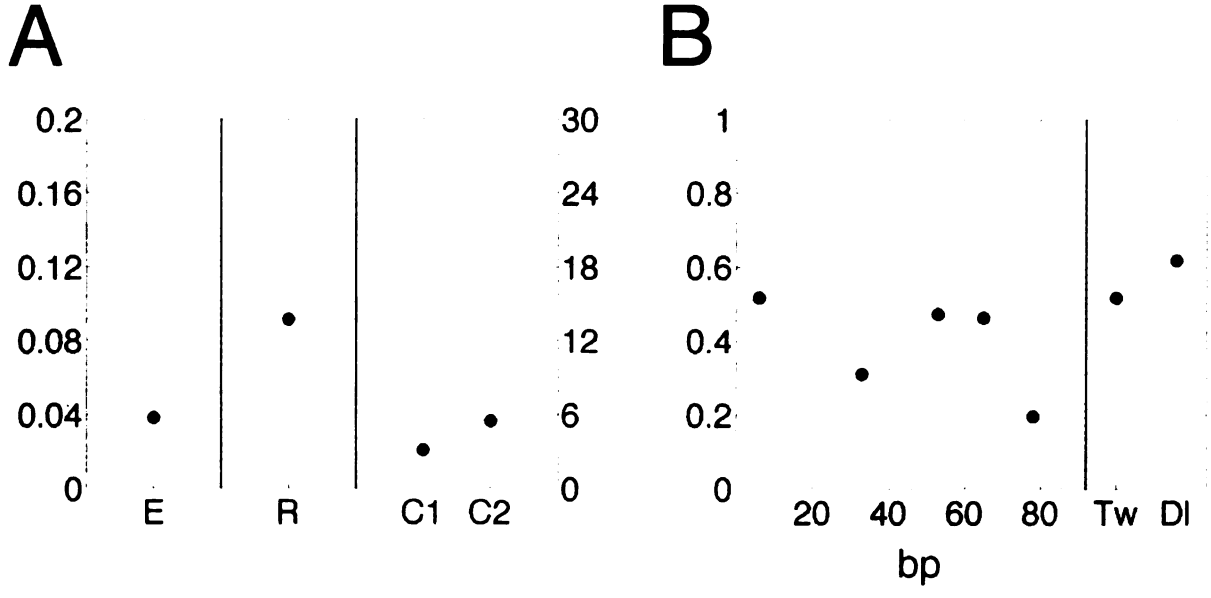
Second, previous qualitative observations show that the effect of short-range repressors decreases with distance, and is lost around 100-150bp. To our knowledge, our study is the first that analyzes distance dependency of the short-range repressors systematically. Short-range repressor quenching efficiency is represented by several parameters in the model as described previously.

We noted a general decrease in quenching efficiency with distance, consistent with previous qualitative observations, but at (52-55) bp, relative efficiency is predicted to increase, before dropping off with greater distance (Figure III-4B). This trend was evident for multiple formulations of the model (Appendix A Figure 5), and persisted when we carried out parameter estimation with subsets of the data (see below), indicating that the non-monotonic behavior reflects a real biochemical property of the Giant repressor. The change in this monotonic behavior may be a reflection of specific phasing effects, perhaps relating to nucleosomal structure. The non-monotonic decline in repression effectiveness was an unexpected result of our modeling and contrary to the simple step functions or linear functions used in previous modeling efforts (Janssens *et al*, 2006; Zinzen *et al*, 2006). Note that the reduction in repression efficiency at ~30bp does not imply that gene 3 (2Gt.35.2Tw.2Dl) should have weak repression, because this gene has an additional more distal binding site that also contributes to activity through quenching and cooperativity.

Third, the repressor quenching efficiency parameters are similar whether the repressor was located adjacent to the Twist or to the Dorsal activator site, which suggests that short-range repressors have similar effects on different activators (Figure III-4B). The short-range repression mechanism appears to involve chromatin modification, which

Figure III-4: Parameters found by the ES parameter estimation technique for scheme 2 of the model.

(A) Root mean square error, E, is shown on the left, with corresponding scale shown on the left axis. Repressor scaling factor R (referred to as S_R in fractional occupancy model in Materials and Methods) and cooperativity C are shown in the central and right portions respectively, with scale shown on the right axis. (B) Quenching efficiency parameters are shown for increasing distances of repressors located 5' of the activators on the left. Quenching efficiency levels relative to Twist proximal (T) sites and Dorsal proximal (D) sites are shown in the right panel. A non-monotonic decrease in quenching efficiency for increasing distances is observed.



may allow for more promiscuous action on many types of transcription factors, rather than a mechanism based on specific contacts between repressor and activator (Li Li, unpublished data). This activator insensitivity is consistent with the action of short-range repressors on a range of enhancers that bind diverse transcriptional enhancers (Gray *et al*, 1994; Kulkarni & Arnosti, 2005). Parameters identified in this study are therefore likely to be generally applicable to diverse settings.

We tested whether the nonlinear quenching is critical to obtaining reasonable parameters by repeating our procedure with a constraint that required a monotonic decrease for quenching efficiency. As shown in Figure III-5, this constraint produced parameter sets that predicted repressor quenching efficiency would remain almost constant between 6bp-77bp, which is not supported by this or previous studies. For example, the 35bp increase in spacing between gene 2 and gene 5 has a measurable effect. Therefore the non-monotonic decrease in quenching efficiency is likely to indicate some actual biological property of the repressors and should be validated experimentally.

The recent fractional occupancy modeling of 44 endogenous *Drosophila* enhancers identified potential cooperativity values that were somewhat greater than those found here. We ran our parameter estimation algorithm with fixed Giant cooperativity values found in Segal *et al* (2008) and estimated the remaining parameter values in our model. We observed that although the main conclusions of our study did not change, the overall fitting was slightly worse (Figure III-6). We extended this analysis by running our parameter estimation algorithm with eight more choices of Giant cooperativity values. Although we tested Giant cooperativity values ranging from 0 to 30, the root mean square

Table III-1: Parameter descriptions for Scheme 2.

In the first column, 12 synthetic enhancers used for parameter estimation in this study are listed. In the second column, parameter selections are shown. In the third column structure of the synthetic enhancers are depicted.

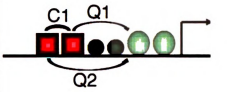
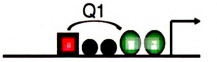
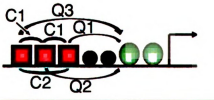
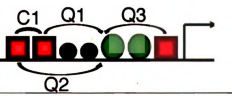
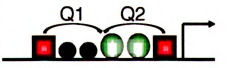
Gene	Parameter Assignments	Gene Structure
1. 2Gt.2Tw.2DI	Q1=q1 Q2=q2	
2. 2Gt.25.2Tw.2DI	Q1=q2 Q2=q3	
3. 2Gt.35.2Tw.2DI	Q1=q2 Q2=q4	
4. 2Gt.50.2Tw.2DI	Q1=q3 Q2=q5	
5. 2Gt.60.2Tw.2DI	Q1=q4 Q2=0	
9. 1Gt.2Tw.2DI	Q1=q1	
16. 1Gt.25.2Tw.2DI	Q1=q2	
10. 3Gt.2Tw.2DI	Q1=q1 Q2=q2 Q3=q3	
17. 3Gt.50.2Tw.2DI	Q1=q3 Q2=q5 Q3=0	
12. 2Gt.2Tw.2DI.1Gt	Q1=q1 Q2=q2 Q3=q6	
14. 1Gt.2Tw.2DI.1Gt	Q1=q1 Q2=q6	
19. 1Gt.50.2Tw.2DI.1Gt	Q1=q3 Q2=q6	

Figure III-5: Parameters for scheme 2 with the constraint that quenching efficiency parameters decrease monotonically.

(A) Root mean square error E, repressor scaling factor R, and cooperativity C labeled as in Figure III-4. (B) Quenching efficiency parameters and relative quenching of Dorsal and Twist sites. Under this constraint, the level of quenching efficiency changes very little from 28-66bp, in contrast to observed trends (Figure III-2).

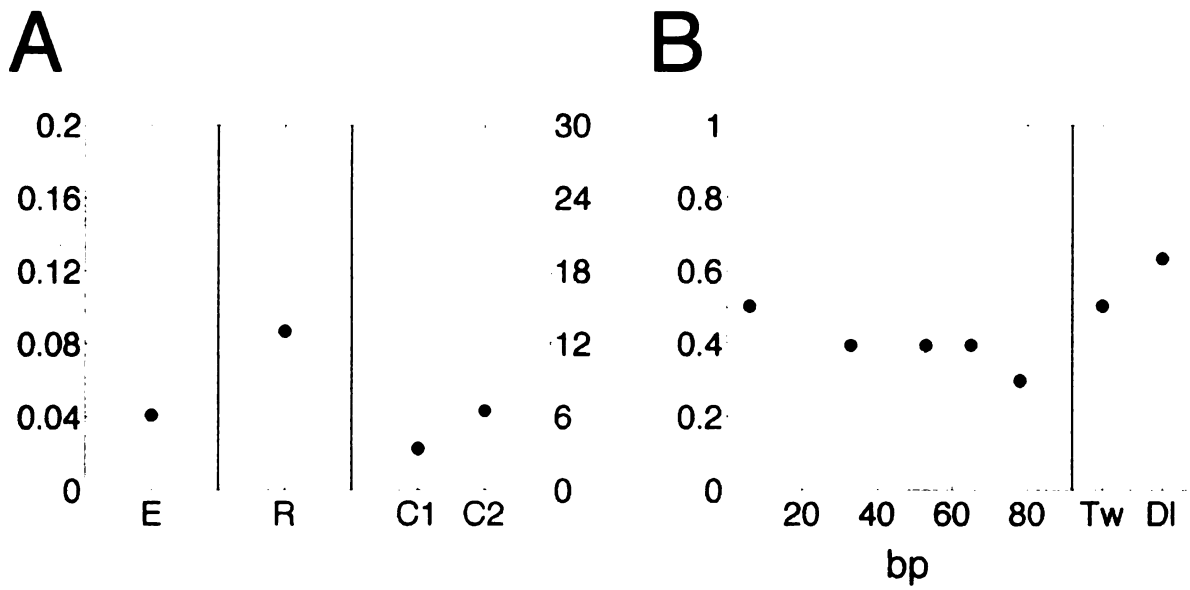
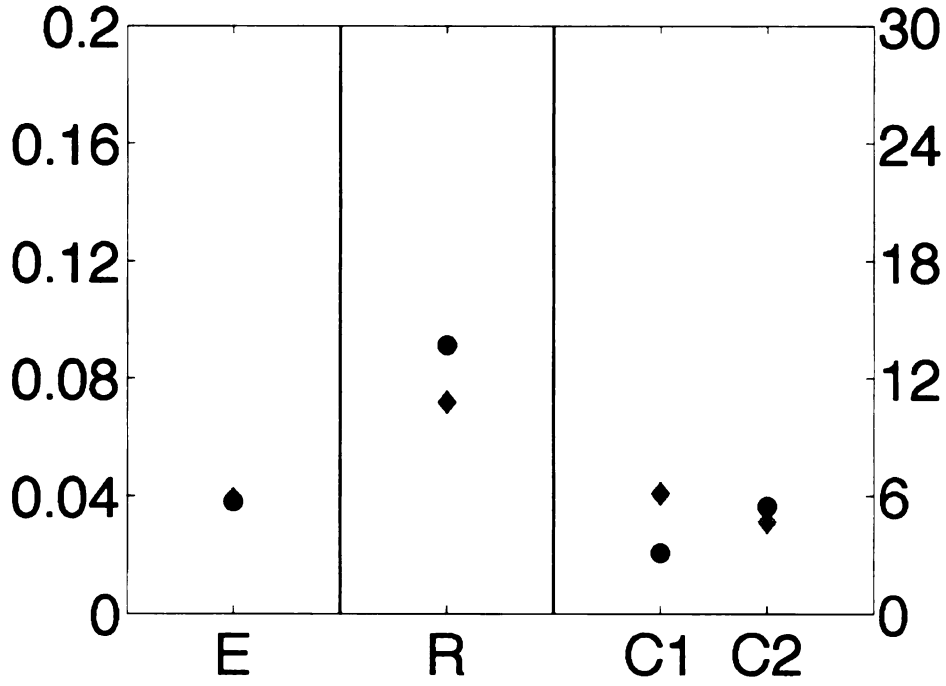


Figure III-6: Parameters for scheme 2 with cooperativity parameters set to different levels. (A, B) Parameters found in our study (circles) and parameters found by constraint of cooperativity parameters to those from in Segal *et al* (2008) (diamonds). The increased cooperativity value is compensated by a decreased repressor scaling factor R. (C) Root mean square errors (RMSE) for cooperativity parameters (constrained to values between 0-30). Estimated cooperativity values from our model lie near the lowest point in this curve.

A


0.2
0.16
0.12
0.08
0.04
0
30
24
18
12
6
0
E
R
C1
C2

B

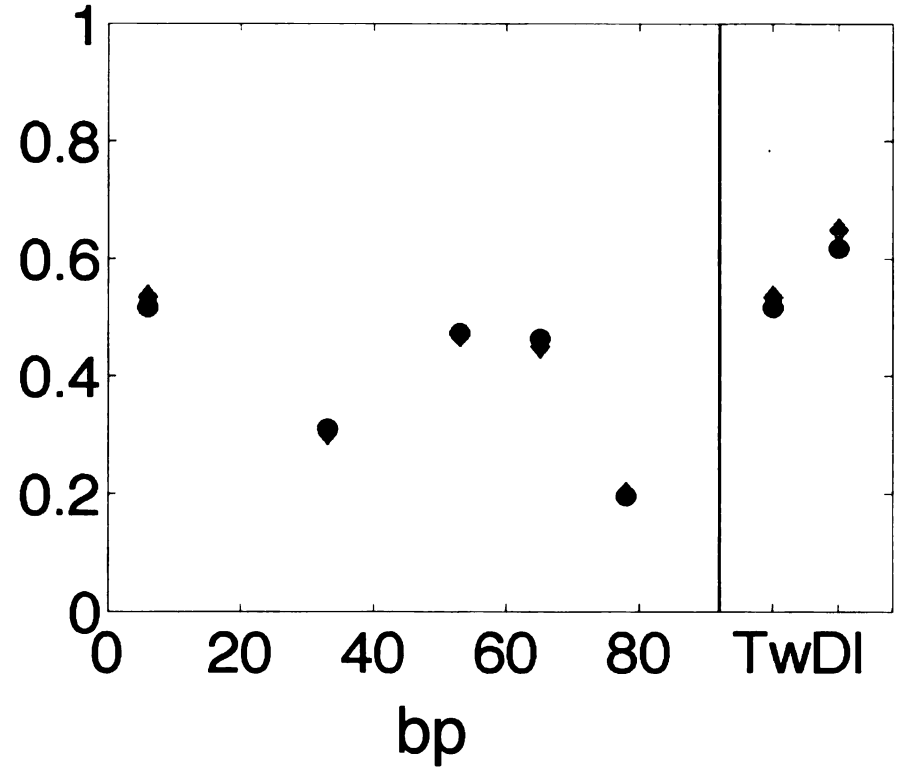
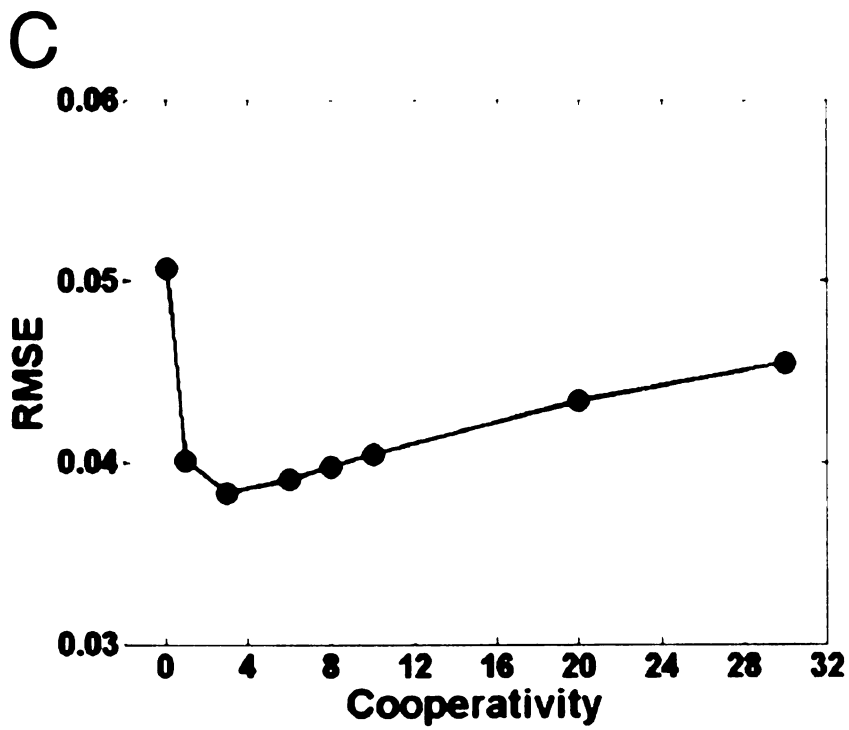

1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
TwDI
bp

Figure III-6: Continued.



errors between predicted and observed values did not change drastically, with minimum at cooperativity value 3.

Model Validation

The analysis described above involved identifying parameters using all data available. An important question is whether such values are "over fit", and whether the model and parameter estimation technique are robust, i.e. relatively insensitive to contributions of individual portions of the data set. Robustness of the parameter estimation technique was described above; here we

assess the model's effectiveness at predicting subsets of the data. We tested whether parameter estimation was markedly affected by removal of individual genes from the data set ("leave-one-out" analysis) (Figure III-7A). We employed nine different forms of the model to evaluate the effects including different assumptions of cooperativity and quenching. We calculated the average of twelve leave-one-out prediction root mean square errors for each scheme, and used these error values for comparison of schemes (Pizarro *et al*, 2000). As judged by the error comparison, schemes 1-4, 8 and 9 work better than schemes 5-7 in this data set, probably due to the smaller number of parameters (Appendix A Figure 6). Leave-one-out analysis was extended by excluding nine separate, specific groups of genes that share structural properties (Figure III-7B). The sets used for this analysis are described in Table III-2. The results of excluding individual genes or sets of related genes suggest that genes that depend on fewer parameters, such as 1Gt.2Tw.2Dl (gene 9), which has no contribution by repressor-repressor cooperativity, may not be predicted well in our analysis. Thus, the contributions of certain classes of

gene can be great. Parameters found by leave-one-out analysis did not change much, but the parameters found by leaving out specific sets of genes changed depending on the genes chosen (Figure III-7A and B). The predictions for genes 1, 10 and 12 by the parameters estimated from set 8, which excludes genes 1, 10 and 12, are shown in Figure III-7C. We conclude that the set of gene modules tested here adequately sample enhancer design to identify critical elements for repressor activity in a robust manner.

Each embryo, with its thousands of imaged nuclei representing different levels of transcription factors, provides a matrix of input and output values that should in theory suffice to describe the response of a gene construct. However, variations in embryo age, staining, and orientation necessitate multiple images for each gene. We obtained between 30 and 53 good quality images for each gene used in our parameter estimation. To test whether this data set is sufficient, or additional individual images would significantly change the conclusions reached, we sampled randomly 50% or 75% of the images from each reporter gene construct, and repeated the parameter estimation. Reducing the data set by one quarter or even one half does not change the value of estimated parameters drastically or the main conclusions of the paper (Figure III-7D). This result suggests that our data set is sufficiently complete, allowing us to draw significant conclusions. In contrast, as shown above, decreasing the number of genes rather than just the number of images obtained for each gene, can affect our results drastically (Figure III-7A and B).

Extension of the model to other repressors and endogenous regulatory elements

Our modeling focused on repression mediated by Giant, which possesses quenching properties similar to those of Snail, Krüppel, and Knirps (Gray *et al*, 1994; Hewitt *et al*, 1999; Kulkarni & Arnosti 2005). To extend these findings to other short-range repressors, Krüppel and Knirps were tested in parallel genes containing one, two, or three binding sites (genes 21-26). As was evident from qualitative staining, both proteins mediated repression, but Krüppel appeared to be a more effective repressor in terms of completeness of reduction of *lacZ* activity. We measured Knirps or Krüppel protein levels with antibodies as was done with Giant, and created [*lacZ*] vs. [repressor] plots for parameter fitting. The limited number of genes tested for these factors did not exhaustively explore possible architectural features, thus making it difficult to differentiate effects of spacing, cooperativity, and relative activity. We judged distance parameters most likely to be conserved between these different factors, based on previously tested genes; therefore the modeling was carried out using quenching parameters from Giant (Gray *et al*, 1994; Arnosti *et al*, 1996a). Modeling was performed to identify likely scaling factors and cooperativity constants. Using the same form of the model used for Figure III-4, we found that cooperativity parameters were low (e.g. Krüppel = 2; Knirps = 0.67), similar to those observed for Giant (Figure A-5; Table A-V). The major difference between Krüppel and Knirps was the repressor scaling factor, which was low in the case of Knirps (~1.4), and more robust for Krüppel (~30), similar to that of Giant (~14). Differences in repression efficiency may be attributed to distinct levels of cooperativity, but the model suggests that such homotypic interactions are of minor importance. This prediction suggests that the higher effectiveness of Krüppel is likely to due to greater potency of this protein on a molar basis, a more complete

occupancy of the binding sites due to their higher affinity, or higher concentrations of the repressor. Further analysis will be required to separate these effects.

Dorsal and Twist activators were studied previously in the context of the *rhomboid* (*rho*) neuroectodermal enhancer (NEE), where their activity was used to identify properties of short-range repressors, including Snail. This protein is required to block expression of *rho* in the mesoderm, resulting in two lateral stripes of expression in the presumptive neuroectoderm of the blastoderm embryo (Figure III-8A). Four Snail binding sites are located within the 330 bp minimal NEE enhancer, and loss of these sites strongly attenuates repression, permitting expression in the mesoderm (Gray *et al*, 1994). A single Snail site (#2) is sufficient to mediate repression, and similar repression is effected by ectopic Snail, Krüppel, or Knirps sites introduced 5' and 3' of the Dorsal 1 and 4 sites respectively, or even a single Snail site 3' of the Dorsal 4 site (Figure III-8A) (Gray *et al*, 1994; Arnosti *et al*, 1996a).

As an extension of our analysis, we tested quenching parameters produced from our model on this element, and carried out parameter estimation to determine values of cooperativity and scaling factors. This modeling is more complex than that employed for genes in Figure III-2, because we now consider scaling factors for each transcription factor, not just for the repressor, and binding sites of different qualities are considered. Position weight matrix (PWM) information was used to score Dorsal, Twist, and Snail sites within the *rho* NEE. In addition, we consider cooperativity not just between repressor sites, but also between activator sites, both of heterotypic (Dorsal-Twist) and homotypic (Twist-Twist) nature. A further consideration is that information about these

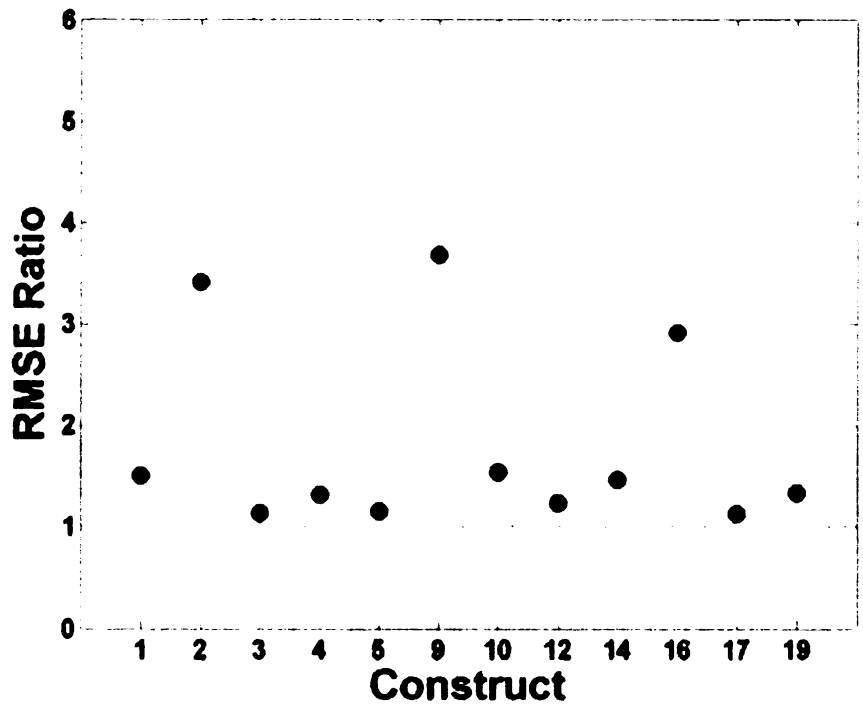
Table III-2: Functionally grouped sets of gene constructs, used for leave-sets-out analysis shown in Figure III-7B.

Set #	Excluded Genes
1	Genes with one or three Giant binding sites (9, 10, 12, 16 and 17)
2	Stoichiometry Genes (1, 9 and 10)
3	Genes with adjacent Giant binding sites in both 5' and 3' end of activators (12, 14 and 19)
4	Genes with only one Giant binding site (9 and 16)
5	Genes with exactly three Giant binding sites (10, 12 and 17)
6	Genes with one Giant binding site at 5' end of activators (9, 14, 16 and 19)
7	Genes with one Giant binding site adjacent to the 5' end of activators (9 and 14)
8	Genes with at least two Giant binding sites adjacent to the 5' end of activators (1, 10 and 12)
9	Genes with three Giant binding sites adjacent to the 5' end of activators (10 and 17)

Figure III-7: Validation of modeling by prediction of subsets of the data from parameters derived from the remainder of the data.

(A) Leave-one-out analysis. Root mean square errors are calculated using parameters found by 11 genes excepting the genes indicated, and all the genes. Relative RMSE ratios, indicating greater errors for prediction of genes 2, 9 and 16, indicating their greater contribution to the parameter constraints. (B) Leave-sets-out analysis for nine distinct sets of genes defined by their shared properties (Table III-2). Root mean square errors are calculated using parameters found from the reduced set and the entire set. Relative RMSE ratios, indicating greater errors for prediction of sets 1, 2 and 4, indicating their greater contribution to the parameter constraints. (C) Predictions for leaving out set 8. Gene 1, 10 and 12 are predicted by using parameters found from other 9 genes. Points represent average values for [*lacZ*] vs. [Gt] data which was divided into 20 bins. (D) Parameter estimation results are shown for different amounts of data 50%, 75% and 100%. The data is cut randomly from each gene at the same percentage.

A


RMSE Ratio
0
1
2
3
4
5
6
1
2
3
4
5
9
10
12
14
16
17
19
Construct

B

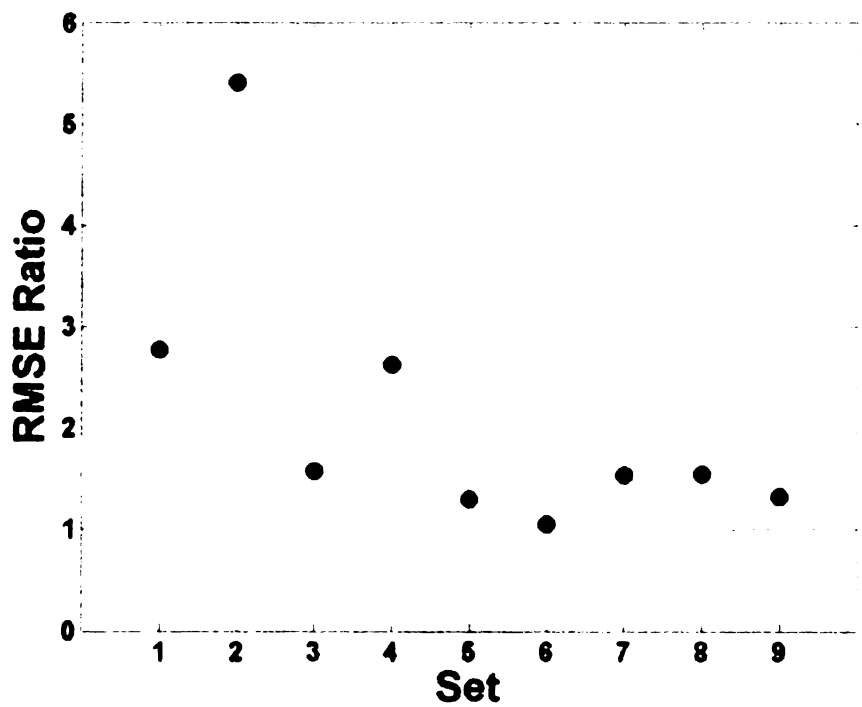

RMSE Ratio
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
Set

Figure III-7: Continued.

C

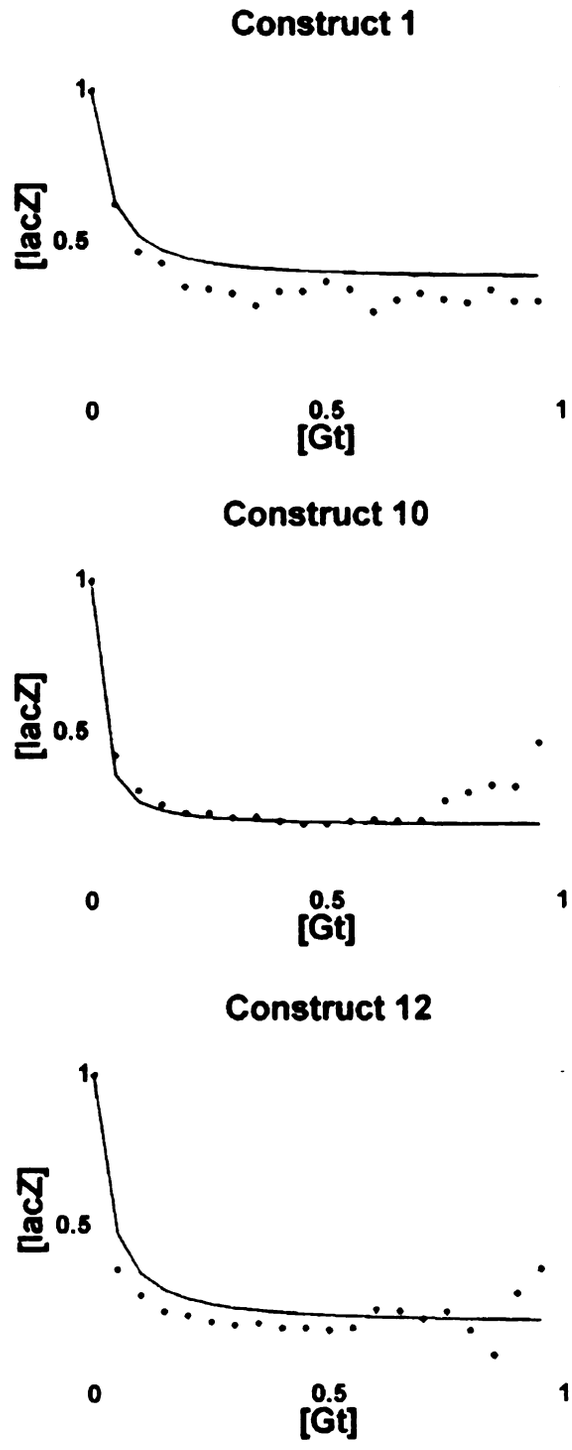
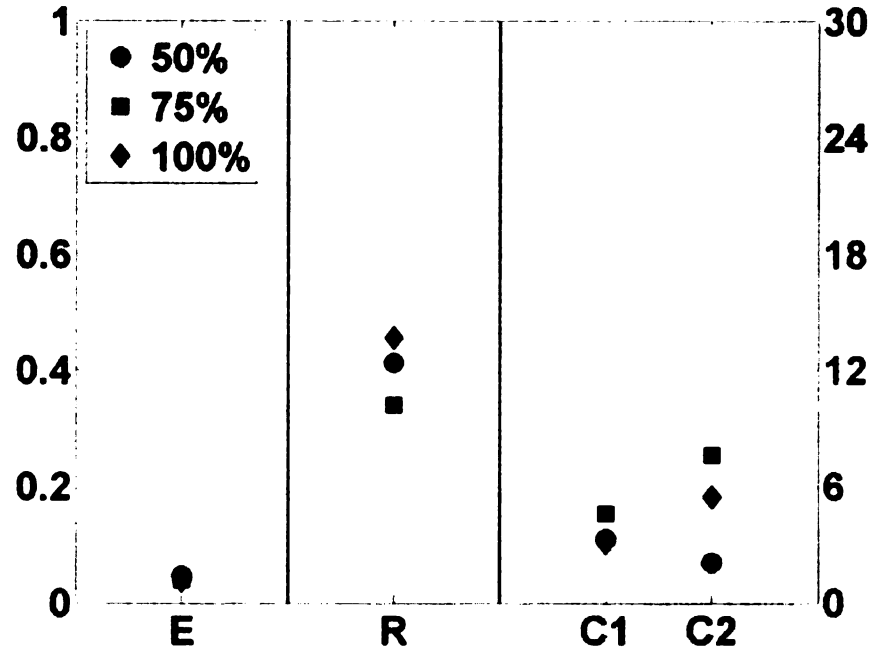
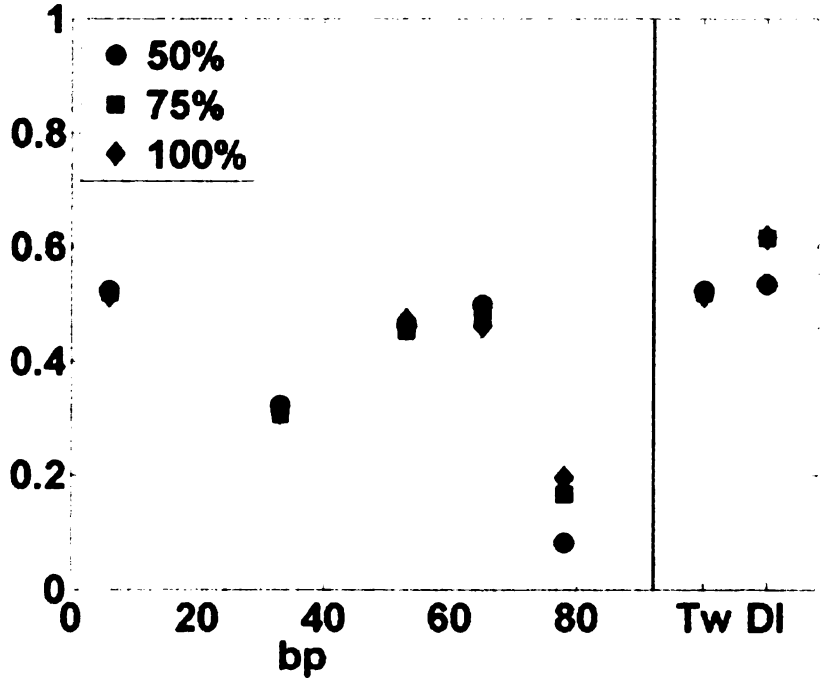


Figure III-7: Continued.

D





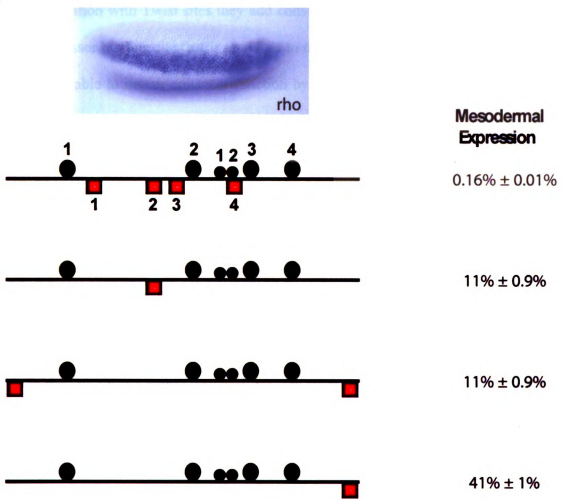
rho NEE variants is qualitative; a single Snail site can repress but may not be as effective as four Snail binding sites.

Simultaneous parameter estimation was carried out using the forms of the *rho* NEE shown in Figure III-8A. We estimated levels of mesodermal repression to be greater than 90% for the endogenous gene, 70-90% for genes carrying one Snail #2 binding site or two ectopic binding sites located 5' and 3' of the element, and 50-70% for one Snail site located 3' of Dorsal #4. Evolutionary strategy parameter estimation was performed multiple times to identify parameters for cooperativity and scaling factors, as well as the predicted effect on expression within ranges specified above. Several striking outcomes were evident from this exercise; first, to find optimal values, the model consistently predicts that the wild-type *rho* NEE, containing four Snail sites, will have output at the lowest end of the allowed range, close to zero, while the internal Snail site #2, or the two ectopic flanking Snail sites, generate values close to the bottom of the allowed range, at about 10% residual activity (Figure III-8A). The single ectopic Snail site 3' of Dorsal #4 is predicted to mediate repression in the middle of the allowable range, about 40% residual activity, consistent with published images (Gray *et al*, 1994). The scaling factor for Dorsal (i.e. its overall activity) is considerably lower than that predicted for Twist, while the scaling factor for Snail is similar to those of Krüppel and Giant (Figure III-8B). Dorsal-Twist cooperativity values vary considerably, with Dorsal2-Twist1 cooperativity predicted to be lower than Twist2-Dorsal3, consistent with the closer spacing of the latter two factors (Crocker *et al*, 2008). Twist-Twist cooperativity is also predicted to be high. These relative differences in activator scaling factors and cooperativity values support known features of the *rho* NEE; the low scaling factors for Dorsal sites are consistent

Figure III-8: Extension of the model to endogenous regulatory elements.

(A) The *rhomboid* gene is expressed in the blastoderm embryo in two lateral stripes (one shown in focal plane), under control of the Dorsal and Twist activators. Ventral expression is inhibited by the Snail short-range repressor, which is expressed in the presumptive mesoderm. The *cis*-regulatory modules used for analysis are shown. Different forms of *rhomboid* NEE enhancer are depicted, with varying number and arrangements of Snail short-range repressor binding sites. Dorsal and Twist activators are shown by large and small green circles respectively, and Snail repressors are shown by red squares. On the right are the predicted repression levels caused by Snail binding sites shown in each module based on parameter estimation using this group of enhancers. (B) Predicted parameters for scaling factors for each transcription factor and cooperativity. Average and standard deviation for twenty estimation runs are shown.

A



B

Parameters

Scaling Factors
Dorsal : 1.2 ± 0.13
Twist : 75 ± 18
Snail : 54 ± 6.5

Cooperativity
Dorsal2-Twist1 : 7 ± 1.3
Dorsal3-Twist2 : 74 ± 16
Twist1-Twist2 : 69 ± 22
Snail2-Snail3 : 65 ± 20

with the inability of individual Dorsal sites to mediate robust activation (Ip *et al*, 1992) ; but in combination with Twist sites they add considerably to the output of the enhancer. A single repressor binding site that is not close to most of the activator sites would in this model still be able to impair enhancer function by initiating a chain-reaction collapse of cooperative interactions. The native *rho* NEE does not appear to rely solely on this mechanism, as most of the identified activator sites lie within a short distance of one of the four Snail sites, suggesting a "belt-and-suspenders" redundant approach to repression. It will be interesting to survey the entire set of enhancers targeted by short-range repressors to determine if this feature is consistently observed in most elements.

DISCUSSION

In the last twenty years, essential features of the biochemistry of gene regulation have come into focus, serving to highlight the considerable complexity and multifarious activities of such *cis*-regulatory elements. We still lack a comprehensive picture of how a transcriptional enhancer operates however. Quantitative models, based on aspects of the system that are readily quantifiable, such as primary DNA structure of a regulatory region, quantities of regulatory proteins, and transcript levels, appear to offer an alternative route to learn about key features of regulatory systems. When combined with biochemical and genomic information, such models may provide the "bridge" that will allow deeper understanding of the function and evolution of *cis* regulatory elements, which are the nexus of many biological processes.

In this study, by employing a reductionist analysis of short-range repression, we sought to explore a relatively untouched, yet central aspect of gene regulation in

Drosophila. Previous qualitative studies highlighted the extreme distance-dependence of short-range repressors, and comparative analysis has shown many instances of evolutionary plasticity of regulatory regions controlled by these proteins (Gray *et al*, 1994; Ludwig & Kreitman, 1995; Hewitt *et al*, 1999; Hare *et al*, 2008). Knowing that transcription factors influence each other in a local fashion permitted the identification of novel enhancers, based on the clustering of binding sites (Berman *et al*, 2002; Schroeder *et al*, 2004). Yet these studies alone do not provide the basis for predicting evolutionary changes that reshape transcriptional output, or predicting comparative activity of enhancers controlled by similar groups of transcription factors. For example, the original hypothesis that the affinity and or number of Bicoid binding sites dictates the output of regulated genes has been replaced by an understanding that other, as-yet unknown features, appear to play more decisive roles (Driever *et al*, 1989; Gao *et al*, 1996; Ochoa-Espinosa *et al*, 2009).

Previous modeling studies focused on endogenous enhancers, which have complex arrangements of transcription factor binding sites, thus even relatively subtle changes made to these elements potentially influence a number of factors. This complexity required an oversimplified treatment of quenching (arbitrary assumption of linear decreases) and cooperativity (only homotypic interactions considered) (Janssens *et al*, 2006; Segal *et al*, 2008). Alternatively, one study considered simplified regulatory elements, but only modeled them in silico, without the experimental treatment we use here (Zinzen *et al*, 2006). It is not clear that the earlier studies are sufficiently robust to predict why extensively reshuffled enhancer sequences may retain similar function in some instances, or show quantitatively distinct outputs in others (Ludwig *et al*, 2005;

Crocker *et al*, 2008). Our studies are by their design shaped to detect quantitative differences resulting from minor differences in repressor-activator spacing, for example, and should be useful for such evolutionary studies. We limited the number of features that differed between elements, allowing modeling with a tractable number of parameters. We utilized a common block of Dorsal and Twist activator sites, allowing us to focus on changes made in the number and arrangement of repressor sites; clearly, differences in affinity, number and arrangement of activator sites also play decisive roles in dictating transcriptional output, thus future modeling efforts will need to integrate these elements as well. The tight focus on short-range repressors with the analysis of a relatively small number of reporter genes provided sufficient data for robust estimation of key parameters (Figure III-7). From our comparison of repression by other short-range repressors, it is likely that the analysis of Giant can guide studies of other similarly-acting repressors, including Krüppel, Knirps and Snail (Figure III-8).

Relating to transcriptional grammar, our study uncovered specific quantitative features that appear to apply to short-range repressors in a general context. We identified a complex nonlinear quenching relationship that suggests that within the range of activity, Giant, and probably other short-range repressors, have an optimum distance of action that may reflect steric constraints (Figure III-4). Multiple formulations of the model generated very similar predictions, suggesting that this nonlinear distance function is a real feature of the system (Appendix A Figure III-5). Consistent with this notion, a previous study of transcription factor binding sites in *Drosophila* enhancers discovered an overall preference of Krüppel sites to be found 17 bp from Bicoid activator sites,

which may be an indication that other short-range repressors also have preferred distances for optimal activity (Makeev *et al*, 2003).

The similar quenching efficiencies for repressors acting adjacent to Dorsal or Twist activator sites was an additional significant finding (Figure III-4). The similar effect on disparate activator proteins indicates that the effects of short-range repression are general, and are likely to be translatable to distinct contexts. Previous empirical tests had already pointed in this direction; for example, insertion of ectopic binding sites for Knirps and Krüppel into *rho* NEE sequences is sufficient to induce repression, although these proteins do not usually cross regulate (Gray *et al*, 1994; Arnosti *et al*, 1996a). In addition, systematic comparisons suggest that short-range repressors can counteract a variety of transcriptional activation domains with similar efficiency, suggesting that specific protein-protein contacts are not essential (Arnosti *et al*, 1996b; Kulkarni & Arnosti 2005). This flexibility plays well into a computational understanding of transcriptional information in the fly, because it is likely that the quantitative correlations we observe in specific genes tested here are general properties of the repressors. The application of our modeling to the *rho* NEE indicates that certain features can be directly applied to more general contexts (Figure III-8).

In one area we found quantitative differences between parameters derived from the synthetic gene modules and the endogenous regulatory regions. The importance of homotypic cooperativity predicted for Snail sites in the context of the *rho* NEE was overall much higher than that found for Giant, Krüppel and Knirps sites acting on the synthetic gene constructs; this might be an example where the individual proteins do exhibit different context dependencies perhaps because the proteins differ in level of

"stickiness". Alternatively, the distance between the Snail sites in question, 23 bp, might facilitate cooperative interactions much more than the closely apposed spacing used in our genes, where steric interference may play an opposing role.

In revisiting previous incarnations of the endogenous *rho* NEE, we highlighted several features of the architecture of this regulatory region. This enhancer appears to be "over-engineered" in terms of the use of Snail to mediate repression; based on previous experiments, it appears that even a single Snail site is sufficient to mediate repression (Gray *et al*, 1994). Consistent with this view, our short-range repressor parameters would predict that for many values of the Snail repressor scaling factor, the enhancer should be effectively inhibited by the set of endogenous binding sites (Figure III-8). How might we understand this apparent redundancy? Such design may provide the correct dynamical response, with a swift repression of *rho* at an early enough time where Snail levels are still low, or it may ensure that gene output is robust to environmental and genetic noise. To carry out this modeling, we made explicit assumptions about the relative repression effectiveness of different configurations of this enhancer, containing one, two or four Snail sites. These different thresholds of repression seem to be warranted, but it would be clearly desirable to have quantitative values for the relative expression, similar to those obtained for our synthetic gene modules, to better model the activity of these enhancers.

The *rho NEE* modeling also highlighted features of transcriptional activators. Activator scaling factors for Dorsal were reproducibly lower than those of Twist, and this was apparent for several different assumptions of expression level (Figure III-8 and data not shown). The relative differences in contribution to activation can be explained by examination of the structure of the enhancer; contribution by the low intrinsic values of

Dorsal are amplified by strong cooperativity with Twist, setting up a chain of interacting weak sites that together are highly active. Experimental evidence bears out these conclusions; isolated Dorsal sites tested on reporter genes mediate relatively weak activation, and a *rho* NEE lacking Twist sites, but containing four Dorsal sites, is similarly compromised (Ip et al, 1992; Szymanski & Levine, 1995).

Our earlier study have suggested that many developmental enhancers, including those regulated by short-range repressors, may possess a flexible "billboard" design, in which individual factors or small groups of proteins would independently communicate with the promoter region, so that the net output of an enhancer would reflect the cumulative set of contacts over a short time period (Kulkarni & Arnosti, 2003). Such a view of enhancers would account for the evolutionary plasticity observed in regulatory sequences. No DNA-scaffolded superstructure, reflecting the formation of a unique three dimensional complex, would be necessary in this scenario. Yet our modeling suggests that the *rho* NEE might involve communication between relatively distant binding sites, through sets of cooperative interactions. In this case, it is possible that such distant interactions might be compatible with a flexible structure, if many distinct configurations of binding sites provide such a cooperative network. Current studies have indeed highlighted potential frameworks involving Dorsal and interacting factors on same classes of enhancer (Erives & Levine, 2004; Papatsenko & Levine, 2007). Application of a transcriptional grammar integrating activities of activators and repressors is a critical next step to illuminate enhancer design and evolution.

REFERENCES

Ackers GK, Johnson AD, Shea MA (1982) Quantitative model for gene regulation by λ phage repressor. *Proc Natl Acad Sci USA* **79:** 1129-1133

Arnosti DN, Gray S, Barolo S, Zhou J, Levine M (1996a) The gap protein knirps mediates both quenching and direct repression in the Drosophila embryo. *EMBO J* **15:** 3659-3666

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

Ay A, Fakhouri WD, Chiu C, Arnosti DN (2008) Image processing and analysis for quantifying gene expression from early Drosophila embryos. *Tissue Eng Part A* **14:** 1517-1526

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci USA* **2:** 757-762

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R (2005a) Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* **15:** 116-124

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R (2005b) Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev* **15:** 125-135

Carroll SB, Grenier JK, Weatherbee SD (2001) *From DNA to Diversity*. Blackwell Science, Malden, Massachusetts, USA

Clyde DE, Corado MS, Wu X, Paré A, Papatsenko D, Small S (2003) A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. *Nature* **426:** 849-853

Crocker J, Tamori Y, Erives A (2008) Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol* **6:** e263

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

Erives A, Levine M (2004) Coordinate enhancers share common organizational features in the Drosophila genome. *Proc Natl Acad Sci USA* **101:** 3851-3856

Fomekong-Nanfack Y, Kaandorp JA, Blom J (2007) Efficient parameter estimation for spatio-temporal models of pattern formation: case study of Drosophila melanogaster. *Bioinformatics* **23:** 3356-3363

Gao Q, Wang Y, Finkelstein R (1996) Orthodenticle regulation during embryonic head development in Drosophila. *Mech Dev* **56:** 3-15

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

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

Hewitt GF, Strunk BS, Margulies C, Priputin T, Wang XD, Amey R, Pabst BA, Kosman D, Reinitz J, Arnosti DN (1999) Transcriptional repression by the Drosophila giant protein: cis element positioning provides an alternative means of interpreting an effector gradient. *Development* **126:** 1201-1210

Ip YT, Park RE, Kosman D, Bier E, Levine M (1992) The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the Drosophila embryo. *Genes Dev* **6:** 1728-1739

Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, Kozlov KN, Manu, Myasnikova E, Vanario-Alonso CE, Samsonova M, Sharp DH, Reinitz J (2004) Dynamic control of positional information in the early Drosophila blastoderm. *Nature* **430:** 368-371

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

Kulkarni MM, Arnosti DN (2003) Information display by transcriptional enhancers. *Development* **130:** 6569-6575

Kulkarni MM, Arnosti DN (2005) Cis-regulatory logic of short-range transcriptional repression in Drosophila. *Mol Cell Biol* **9:** 3411-3420

Ludwig MZ, Kreitman M (1995) Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. *Mol Biol Evol* **12:** 1002-1011

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M (2005) Functional evolution of a cis-regulatory module. *PLoS Biol* **3:** e93

Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA (2003) Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res* **31:** 6016-6026

Mayo AE, Setty Y, Shavit S, Zaslaver A, Alon U (2006) Plasticity of the cis-regulatory input function of a gene. *PLoS Biol* **4:** e45

Mendes P, Kell D (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* **14:** 869-883

Moles CG, Mendes P, Banga JR (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res* **13:** 2467-2474

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

Papatsenko D, Levine M (2007) A rationale for the enhanceosome and other evolutionarily constrained enhancers. *Curr Biol* **17:** 955-957

Pizarro J, Guerrero E, Galindo, PL (2000) A statistical model selection strategy applied to neural networks. *Proceedings of the ESANN* **2000:** 55-60

Runarsson TP, Yao X (2005) Search Biases in Constrained Evolutionary Optimization. *IEEE Transactions on Systems, Man and Cybernetics Part C* **35:** 233-243

Sánchez L, Thieffry D (2001) A logical analysis of the Drosophila gap-gene system. *J Theor Biol* **211:** 115-141

Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U (2004) Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol* **2:** e271

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

Setty Y, Mayo AE, Surette MG, Alon U (2003) Detailed map of a cis-regulatory input function. *Proc Natl Acad Sci USA* **100:** 7702-7707

Shea MA, Ackers GK (1985) The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol* **181:** 211–230

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

Stormo, GD (2000) DNA binding sites: representation and discovery. *Bioinformatics* **16:** 16-23

Szymanski P, Levine M (1995) Multiple modes of dorsal-bHLH transcriptional synergy in the Drosophila embryo. *EMBO J* **14:** 2229-2238

Vilar JM, Leibler S (2003) DNA looping and physical constraints on transcription regulation. *J Mol Biol* **331:** 981-989

Von Hippel PH, Revzin A, Gross CA, Wang AC (1974) Non-specific DNA binding of genome regulating proteins as a biological control mechanism: 1. The lac operon: equilibrium aspects. *Proc Natl Acad Sci USA* **71:** 4808-4812

Wharton KA Jr, Crews ST (1993) CNS midline enhancers of the Drosophila slit and Toll genes. *Mech Dev* **40:** 141-154

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

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

Chapter 4

Conclusions and Future Directions

CONCLUSION

Analysis of gene networks and genomic *cis*-regulatory elements holds the key to understanding how genomes encode the properties of organisms. However, despite intensive research, our understanding of the design principles underlying complex genetic switches is currently at a rudimentary level. In the last twenty years, essential features of the biochemistry of gene regulation have come into focus. For the purposes of developing a comprehensive picture of how a transcriptional enhancer operates, however, such studies have served more to highlight the considerable complexity and multifarious activities of such *cis* regulatory elements, rather than to provide a quantitative basis for understanding their activity. Quantitative models, based on aspects of the system that are readily quantifiable, such as primary DNA structure of a regulatory region, quantities of regulatory proteins, and transcript levels, appear to offer an alternative route to learn about key features of regulatory systems. When combined with biochemical and genomic information, such models may provide the "bridge" that will allow deeper understanding of the function and evolution of *cis* regulatory elements, which are the nexus of many biological processes.

In this work, we describe studies including image processing, data analysis and mathematical modeling that shed light on how transcriptional regulatory information encoded within the *cis*-regulatory enhancer sequence is interpreted by the cellular machinery. In the image processing and data analysis section (Chapter 2) we describe a method to correlate the transcription factors and mRNA output for gene modeling

purposes. We show that the levels of *lacZ* mRNA, and potentially the transcriptional repressor protein Giant, are proportional to fluorescent intensities, a critical basis for quantitative modeling. Our analysis also reveals that a suggested parabolic form of the background fluorescence in confocal images of early *Drosophila* embryos is evident most prominently in flattened specimens, with intact embryos exhibiting a more linear background. After appropriate background subtraction and normalization, these data are amenable to representation of gene regulatory surfaces that permit creation and validation of quantitative models of gene expression. In this way, we have constructed the foundations for modeling a *cis*-regulatory 'grammar' that applies to an important set of transcriptional regulators in the *Drosophila* blastoderm embryo. More generally, the image and data analysis techniques described in this study can be generalized to understand the quantitative function of endogenous genes or gene networks in the *Drosophila* embryo or other well-characterized systems. Such an approach may also prove useful for engineered transcriptional regulatory elements employed for targeted expression of genes in therapeutic settings.

In the mathematical modeling section (Chapter 3), by employing a reductionist analysis of short-range repression, we sought to explore a relatively untouched, yet central aspect of gene expression from *Drosophila*. Previous qualitative studies played a key role in highlighting the extreme distance-dependence of such repressors, and comparative analysis has shown many instances of evolutionary plasticity of regulatory regions controlled by such repressors (Gray *et al*, 1994; Ludwig & Kreitman, 1995; Hewitt *et al*, 1999; Hare *et al*, 2008). Knowing that transcription factors influence each other in a local fashion permitted the identification of novel enhancers, based on the

clustering of binding sites (Berman *et al*, 2002; Schroeder *et al*, 2004). Yet these studies alone do not provide the basis for predicting evolutionary changes that reshape transcriptional output, or predicting comparative activity of enhancers controlled by similar groups of transcription factors. For example, the original hypothesis that the affinity and or number of Bicoid binding sites dictates the output of regulated genes has been replaced by an understanding that other, as-yet unknown features, appear to play more decisive roles (Driever *et al*, 1989; Gao *et al*, 1996; Ochoa-Espinosa *et al*, 2009).

These previous studies largely relied on endogenous enhancers, which have complex arrangements of transcription factor binding sites, thus even relatively discrete changes made to these elements probably perturb a number of factors. Here, we limited the number of features that differed between elements, allowing modeling with a tractable number of parameters. We utilized a common block of Dorsal and Twist activator sites, allowing us to focus on changes made in the number and arrangement of repressor sites; clearly, differences in affinity, number and arrangement of activator sites also play decisive roles in dictating transcriptional output, thus future modeling efforts will need to integrate these elements as well. The tight focus on short-range repressors with the analysis of a relatively small number of reporter constructs provided data in depth sufficient to provide robust information on key parameters (Figure III-7). From our comparison of repression by other short-range repressors, it is likely that the extensive analysis of Giant provides good guidance for the activity of other similarly-acting repressors; modeling of Krüppel, Knirps and Snail appear to be feasible on elements that they are known to interact with (Figure III-8). To extend this type of analysis to other transcription factors, it is likely that separate empirical studies will be essential to

understand the basic elements of their context effects. In one case, the Hairy long-range repressor exerts repressive influences over greater than one kilobase, suggesting that the constraints on placement of Hairy binding sites on target genes may be quite relaxed compared to those applicable to short-range repressors, an issue that warrants further investigation (Barolo & Levine, 1997). Studies of this sort may not unlock a comprehensive atlas of *cis* regulatory relationships for all transcription factors, but by providing in-depth insights on a one-by-one (or a class-by-class) basis, important advances are feasible relating to the activity and evolution of specific classes of genes. To date, similar model gene studies have provided the basis of our qualitative, empirical understanding of transcription, thus attacking a quantitative problem in a discrete fashion, one group of factors at a time, has strong precedents.

Fractional occupancy modeling has been applied to understand *Drosophila* transcriptional regulation in several other studies. Segal and colleagues combined expression data from 44 characterized blastoderm enhancers and 8 trans-acting factors to identify parameters describing cooperativity and factor activity (Segal *et al*, 2008). This exercise was fairly successful, judged by the ability of the obtained parameters to predict an additional set of 26 enhancers that were not included in the modeling. This form of the model did not consider quenching distances, however, (contributions of activators and repressors are merely summed, with no consideration for relative spacing). In light of the extensive use of short-range repression in these enhancers, this simplification must be considered a major shortcoming. The failure of this model to consider heterotypic cooperativity is also likely to account for some of the limitations, as pointed out by these researchers. Reinitz and colleagues focused in on a 1.7 kbp regulatory region driving

expression of the *even-skipped* gene; by considering possible binding sites for Caudal not included in original DNaseI footprinting studies, this model predicted specific features of expression that were subsequently quantitatively validated (Janssens *et al*, 2006). It is not clear how generally applicable to other enhancers the derived parameters were from this model – quenching distances of short-range repressors are estimated to be 1 for distances less than 50 bp with a linear decrease to 0 for distances between 50 bp and 150 bp, for example. In the former study, the modeling might suffer from simplification required in considering a great number of enhancers that differ in a many different ways, while in the latter study, a single complex element is target of focused studies, with the possible limitation that any model would be liable to overfitting (although mutant forms of the enhancer are also considered in the modeling). A third fractional occupancy modeling study considered regulatory elements controlled by the same activators we tested here, namely Dorsal and Twist (Zinzen *et al*, 2006). The overall objective of this study was to understand the relative activity of enhancers of the neuroectoderm that are controlled by Dorsal, Twist and the Snail short-range repressor. In this model, parameters for cooperativity, activity, and quenching were also considered, and a critical importance of Dorsal-Twist cooperativity was identified, although not Twist-Twist cooperativity. The quenching factor for Snail was again not explicitly measured, but was defined as a simple stepwise function equal to 1 for distances less than 100 and 0 for distances greater than 100 bp. A major difference between this study and those of Segal *et al* (2008) and Janssens *et al* (2006) is that the modeling was conducted in silico on synthetic gene modules with different configurations of binding sites, similar to our approach, however the output of these gene modules was never empirically tested. The main conclusions

from parameter estimation are therefore based on which sets of values might be likely to provide informative results. Thus, our study provides a critical feedback between experiment and quantitative modeling that is lacking in this earlier study. It is not clear that the earlier studies are sufficiently robust to predict why extensively reshuffled enhancer sequences from different retain similar function, or, in some cases, show quantitatively distinct outputs (Ludwig *et al*, 2005; Crocker *et al*, 2008). Our studies are by their design shaped to detect quantitative differences resulting from minor differences in repressor-activator spacing, for example, and should be useful for such evolutionary studies.

This study identified specific quantitative features that appear to apply to short-range repressors in a general context. Our study identified a complex nonlinear relationship that suggests that within the range of activity, Giant, and probably other short-range repressors, have an optimum distance of action that may reflect steric constraints (Figure III-4). Multiple formulations of the model generated very similar predictions, suggesting that this nonlinear distance function is a real feature of the system (Figure A-5). Consistent with this notion, a previous study of transcription factor binding sites in *Drosophila* enhancers discovered an overall preference of Krüppel sites to be found 17 bp from Bicoid activator sites, which may be an indication that other short-range repressors also have preferred distances for optimal activity (Makeev *et al*, 2003).

Another significant finding was the similar quenching efficiencies for repressors acting adjacent to Dorsal or Twist activator sites (Figure III-4). The similar effect on disparate activator proteins indicates that the effects of short-range repression are general, and are likely to be translatable to distinct contexts. Previous empirical tests had already

pointed in this direction; for example, insertion of ectopic binding sites for gap gene repressors Giant, Knirps and Krüppel into *rhomboid* NEE sequences is sufficient to induce repression, although these proteins do not usually cross regulate (Gray *et al*, 1994; Arnosti *et al*, 1996a; Hewitt *et al*, 1999). In addition, systematic comparisons suggest that short-range repressors can counteract a variety of transcriptional activation domains with similar efficiency, suggesting that specific protein-protein contacts are not essential (Arnosti *et al*, 1996b; Kulkarni & Arnosti 2005). The generality of repression suggests that this class of transcription factor is able to be readily recruited to novel contexts, a feature that may have played an instrumental role during the extensive remodeling of transcriptional regulatory circuitry associated with the evolution of the derived syncytial blastoderm of *Drosophila*. This flexibility plays well into a computational understanding of transcriptional information in the fly, because it is likely that the quantitative correlations we observe in specific genes tested here are general properties of the repressors. The application of our modeling to the *rho* NEE indicates that certain features can be directly applied to more general contexts (Figure III-8). The prediction that short-range repressors such as Snail and Giant will generally exhibit similar efficiency acting on other activators, in other contexts, will be an important goal for future application of this modeling.

In one area we found quantitative differences between parameters derived from the synthetic gene modules and the endogenous regulatory regions. The importance of homotypic cooperativity predicted for Snail sites in the context of the *rho* NEE was overall much higher than that found for Giant, Krüppel and Knirps sites acting on the synthetic gene constructs; this might be an example where the individual proteins do

exhibit different context dependencies perhaps because the proteins differ in level of "stickiness". Alternatively, the distance between the Snail sites in question, 23 bp, might facilitate cooperative interactions much more than the closely apposed spacing used in our constructs, where steric interference may play an opposing role.

In revisiting previous incarnations of the endogenous *rho* NEE, we highlighted several features of the architecture of this regulatory region. This enhancer appears to be "over-engineered" in terms of the use of Snail to mediate repression; based on previous mutagenesis experiments, it appears that even a single Snail site located within the central enhancer would be sufficient to mediate repression (Gray *et al*, 1994). Consistent with this view, our short-range repressor parameters would predict that for many values of the Snail repressor scaling factor, the enhancer should be effectively inhibited by the set of endogenous binding sites (Figure III-8). How might we understand this apparent redundancy? Such design may provide the correct dynamical response, with a swift repression of *rho* at an early enough time where Snail levels are still low, or it may ensure that gene output is robust to environmental and genetic noise. We made explicit assumptions about the relative repression effectiveness of different configurations of this enhancer, containing one, two or four Snail sites. These different thresholds of repression seem to be warranted, but it would be clearly desirable to have quantitative values for the relative expression, similar to those obtained for our synthetic gene modules, to better model the activity of these enhancers.

A second feature of this modeling provided an insight into transcriptional activators that was not apparent from our synthetic modules, in which a single cluster of Dorsal and Twist sites was used in different contexts. Activator scaling factors for

Dorsal were reproducibly lower than those of Twist, and this was apparent for several different assumptions of expression level (Figure III-8 and data not shown). The relative differences in contribution to activation can be rationalized by examination of the structure of the enhancer; contribution by the low intrinsic values of Dorsal are amplified by strong cooperativity with Twist, setting up a chain of interacting weak sites that together are highly active. Interference of part of this chain of regulatory interactions by Snail would be sufficient to bring the whole edifice crashing down. Experimental evidence bears out these conclusions; isolated Dorsal sites tested on reporter genes mediate relatively weak activation, and a *rho* NEE lacking Twist sites, but containing four Dorsal sites, is similarly compromised. The relative constrained placement of short-range transcriptional repressors in some contexts should contrast with the expected relaxed constraints for long-range repressors, such as Hairy, that are able to mediate repression from distal sites. We expect that evolutionary constraint of binding sites for this latter type of transcriptional regulator should be much lower, a prediction that will require better identification of physiologically active sites on endogenous target genes.

Our earlier study have suggested that many developmental enhancers, including those regulated by short-range repressors, may possess a flexible "billboard" design, in which individual factors or small groups of proteins would independently communicate with the promoter region, so that the net output of an enhancer would reflect the cumulative set of contacts over a short time period (Kulkarni & Arnosti, 2003). Such a view of enhancers would account for the evolutionary plasticity observed in regulatory sequences. No DNA-scaffolded superstructure, reflecting the formation of a unique three dimensional complex, would be necessary in this scenario. Yet our modeling suggests

that the *rho* NEE might involve communication between relatively distant binding sites, through sets of cooperative interactions. In this case, it is possible that such distant interactions might be compatible with a flexible structure, if many distinct configurations of binding sites provide such a cooperative network. Current studies have indeed highlighted potential frameworks involving Dorsal and interacting factors on same classes of enhancer (Erives & Levine, 2004; Papatsenko & Levine, 2007). Application of a transcriptional grammar integrating activities of activators and repressors is a critical next step to illuminate enhancer design and evolution.

FUTURE DIRECTIONS

Comparative sequence analysis, coupled with the development of algorithms to search genomic databases, has provided important tools for the identification of gene regulatory elements at a scale not previously possible, but this approach has been partially successful at finding *cis*-regulatory motifs. In many cases, predicted regulatory elements are non-functional possibly due to enhancer architecture, which are not taken into account usually by the present techniques. It is necessary to factor in features of enhancers such as binding site number, spacing, orientations and order, in order to achieve better predictions in the interpretation of their biological function. Our study (Chapter 3) takes a first step towards providing such a paradigm for early *Drosophila* developmental enhancers that are regulated by the short-range transcriptional repressors. Our findings on enhancer grammar might be incorporated to present bioinformatics techniques, which employ information about putative binding sites and evolutionary conservations, to survey genomic regions. Such incorporation might provide powerful

predictive tools that will be useful for studies of population biology, such as those where distinct gene expression profiles are linked to disease susceptibility and differences in drug metabolism. A major goal now is to identify novel regulatory elements using all sequenced *Drosophila* genomes and understand how evolutionary processes shape endogenous enhancers, focusing on targets of short-range repressors that have not been well characterized with respect to regulatory elements. Some candidates include A-P patterning gene *even-skipped* and D-V patterning gene *rhomboid*, which show conserved function in other *Drosophila* species with some small changes in enhancer structure.

Current approaches for identification of candidate regulatory modules include computational methods that look for clusters of transcription factor binding sites and phylogenetic comparisons that identify evolutionarily conserved sequences (Berman et al, 2002; Markstein and Levine, 2002; Markstein et al., 2002; Bergman and Kreitman, 2001). However these methods give large number of false-positive and false-negative results for enhancer prediction, possibly due to degenerate nature of transcription-factor-binding sites. Thus, in many cases where *cis*-regulatory modules predicted by computational methods appear suitable, something in the arrangement of sites which has shown to be critical in our study renders the *cis*-element non-functional. In order to achieve better predictions and eliminate false-positive and false-negative results, computational methods should include structural features such as parameters for spacing and position of binding motifs within *cis*-elements in the search algorithms. For example, suppose that a computational search for novel *cis*-elements based on clustering of binding sites for transcription factors was used to identify enhancers by Giant. Given the relative number, affinities, spacing and distribution of repressor and activator sites within the

module we would be better able to predict the possibility of Giant regulating that element and how strong that regulation would be. The neurectoderm enhancers (NEEs) such as *rho, vnd* and *sog* active in the mid-blastoderm *Drosophila* embryo provide an excellent test case for this idea. These enhancers are not cross-regulated by Giant, Krüppel, or Knirps, i.e. these short-range repressors' binding sites located within the critical NEE enhancers cannot be of functional importance, by this criterion. Furthermore, we can model the expression driven by NEEs in terms of transcriptional inputs from the expected regulators (Dl, Twi, Sna) as well as the "inappropriate" repressors (Gt, Kr, Kni), and examine the effect of each putative cross-regulating site. Such an analysis will reveal the presence of any inappropriate sites that are allowed in the NEEs because they are located outside of the zone of influence imposed by short-range repressors.

By combining quantitative methods to expression levels of the transcription factors and a downstream target gene over space and time we can build a predictive modeling tool that will allow us not only the identification of novel *cis*- elements that are regulated by the same suite of transcription factors, in this case the short-range repressors, but also to predict its expression pattern. Computational algorithms based on binding site sequence data from transcription factor binding site databases, can be used to scan the genome for clusters of short-range repressor binding sites together with binding sites for other proteins that are known to work with them. Combining datasets from a number of large-scale analyses might be used for prefiltering the set of putative regulatory elements obtained. Bioinformatic predictions of binding sites might be used to build a predicted picture of the configurations of these regulatory elements, and expected output of the endogenous regions might be obtained by fractional occupancy modeling. A

simplified example of how exactly such a predictive tool can be used is as follows: We can identify the putative targets of short-range transcriptional repressors in the *Drosophila melanogaster* genome and predict how these candidate genes will respond over space and time. To understand the molecular function and design of these elements, we can incorporate the identified parameters for short-range repression in the bioinformatic analysis of known regulatory elements in D. *melanogaster* and in related species, in particular the 12 sequenced fly genomes. For example, we can apply our tool to anterior-posterior and dorsal-ventral patterning enhancers, for which no quantitative modeling approach has adequately considered the special nature of short-range repressors up until now. Such an analysis would be a test for the predicted parameters in our study, by which we should be able to accurately predict expression levels. In addition, we can make predictions about the effect of ectopic binding sites introduced artificially into regulatory sequences; and directly validate these by measurement of transgenes. By these analyses we might identify previously hidden features of such elements, such as possible redundancy and extra regulatory sites not found in simple "promoter-bashing" experiments.

Mathematical models such as Boolean models and differential equation models have been used to understand the gap gene regulatory network in early *Drosophila melanogaster* embryos as mentioned in the introduction (Sanchez *et al,* 2001; Jaeger *et al*, 2004a; Jaeger *et al*, 2004b). These models were in agreement with the earlier mutant and reporter studies. In addition, the differential equation model was also in agreement with the available spatial and temporal data of protein levels. However, experimental analysis of this regulatory cascade point to at least one additional level of complexity that

needs to be included in such predictive models of gene regulatory networks, which is a detailed understanding of the *cis*-regulatory substructure and its functional significance as we have defined for the short range transcriptional repressors (Giant, Knirps and Krüppel) in Chapter 3.

REFERENCES

Arnosti DN, Gray S, Barolo S, Zhou J, Levine M (1996a) The gap protein knirps mediates both quenching and direct repression in the Drosophila embryo. *EMBO J* **15:** 3659-3666

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

Barolo S, Levine M (1997). Hairy mediates dominant repression in the Drosophila embryo. *EMBO J.* **16**: 2883-2981

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci USA* **2:** 757-762

Bergman, C.M., Kreitman, M. (2001). Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11(8)**: 1335--1345.

Crocker J, Tamori Y, Erives A (2008) Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol* **6:** e263

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

Erives A, Levine M (2004) Coordinate enhancers share common organizational features in the Drosophila genome. *Proc Natl Acad Sci USA* **101:** 3851-3856

Gao Q, Wang Y, Finkelstein R (1996) Orthodenticle regulation during embryonic head development in Drosophila. *Mech Dev* **56:** 3-15

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

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

Hewitt GF, Strunk BS, Margulies C, Priputin T, Wang XD, Amey R, Pabst BA, Kosman D, Reinitz J, Arnosti DN (1999) Transcriptional repression by the Drosophila giant

protein: cis element positioning provides an alternative means of interpreting an effector gradient. *Development* **126:** 1201-1210

Jaeger J, Blagov M, Kosman D, Kozlov KN, Manu, Myasnikova E, Surkova S, Vanario-Alonso CE, Samsonova M, Sharp DH, Reinitz J (2004) Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster. *Genetics* **167(4):** 1721-1737

Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, Kozlov KN, Manu, Myasnikova E, Vanario-Alonso CE, Samsonova M, Sharp DH, Reinitz J (2004) Dynamic control of positional information in the early Drosophila blastoderm. *Nature* **430(6997):** 368-371

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

Kulkarni MM, Arnosti DN (2003) Information display by transcriptional enhancers. *Development* **130:** 6569-6575

Kulkarni MM, Arnosti DN (2005) Cis-regulatory logic of short-range transcriptional repression in Drosophila. *Mol Cell Biol* **9:** 3411-3420

Ludwig MZ, Kreitman M (1995) Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. *Mol Biol Evol* **12:** 1002-1011

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M (2005) Functional evolution of a cis-regulatory module. *PLoS Biol* **3:** e93

Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA (2003) Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res* **31:** 6016-6026

Markstein M and Levine M. (2002) Decoding cis-regulatory DNAs in the Drosophila genome. *Curr Opin Genet Dev.* **12,** 601-606.

Markstein M, Markstein P, Markstein V, and Levine M. (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *PNAS* **99,** 763-768.

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

Papatsenko D, Levine M (2007) A rationale for the enhanceosome and other evolutionarily constrained enhancers. *Curr Biol* **17:** 955-957

Sánchez L, Thieffry D (2001) A logical analysis of the Drosophila gap-gene system. *J Theor Biol* **211:** 115-141

Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U (2004) Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol* **2:** e271

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

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

APPENDIX A

Supplementary Material for Chapter III

Modeling:

Here we describe a general fractional site occupancy model and then explain how it is applied in our case. Consider an enhancer, with two activator and two repressor binding sites bound by activators A_1 and A_2, and repressors R_1 and R_2 respectively. In our case, the repressors are short-range repressors. Simultaneous binding of the repressors and activators may result in a partially active state of the enhancer.

Different bound state of the enhancer, for example, two activators and one repressor, represent distinct states. The probability of each state of those proteins on the DNA is described by the following parameters.

$[A_1]$ and $[A_2]$: Concentration of activator 1 and 2 respectively,

$[R_1]$ and $[R_2]$: Concentration of repressor 1 and 2 respectively,

K_{A_1} and K_{A_2} : Binding affinity of activator 1 and 2 respectively,

K_{R_1} and K_{R_2} : Binding affinity of repressor 1 and 2 respectively,

C_{12}^{a} and C_{12}^{r}: Cooperativity of activators and repressors for binding to the DNA respectively.

$q_{R_i}^{A_j}$: Quenching efficiency of repressor R_i on the activator A_j.

F_X: Probability of having X case.

For simplification in the following formulas,

$$Z = (1 + K_{A_1}[A_1] + K_{A_2}[A_2] + C_{12}^{a} K_{A_1}[A_1] K_{A_2}[A_2])(1 + K_{R_1}[R_1] + K_{R_2}[R_2] + C_{12}^{r} K_{R_1}[R_1] K_{R_2}[R_2])$$

No activator and repressor bound to the DNA: $F_N = \frac{1}{Z}$

Activator 1 is bound to the DNA: $F_{A_1} = \frac{K_{A_1}[A_1]}{Z}$

Activator 2 is bound to the DNA: $F_{A_2} = \frac{K_{A_2}[A_2]}{Z}$

Repressor 1 is bound to the DNA: $F_{R_1} = \frac{K_{R_1}[R_1]}{Z}$

Repressor 2 is bound to the DNA: $F_{R_2} = \frac{K_{R_2}[R_2]}{Z}$

Activator 1 and 2 are bound to the DNA: $F_{A_1A_2} = \frac{C_{12}^{a} K_{A_1}[A_1] K_{A_2}[A_2]}{Z}$

Activator 1 and repressor 1 are bound to the DNA: $F_{A_1R_1} = \frac{K_{A_1}[A_1] K_{R_1}[R_1]}{Z}$

Activator 1 and repressor 2 are bound to the DNA: $F_{A_1R_2} = \frac{K_{A_1}[A_1] K_{R_2}[R_2]}{Z}$

Activator 2 and repressor 1 are bound to the DNA: $F_{A_2R_1} = \frac{K_{A_2}[A_2] K_{R_1}[R_1]}{Z}$

Activator 2 and repressor 2 are bound to the DNA: $F_{A_2R_2} = \frac{K_{A_2}[A_2] K_{R_2}[R_2]}{Z}$

Repressor 1 and 2 are bound to the DNA: $F_{R_1R_2} = \frac{C_{12}^{r} K_{R_1}[R_1] K_{R_2}[R_2]}{Z}$

Activator 1 and 2, and repressor 1 are bound to the DNA:

$$F_{A_1A_2R_1} = \frac{C_{12}^{a} K_{A_1}[A_1] K_{A_2}[A_2] K_{R_1}[R_1]}{Z}$$

Activator 1 and 2, and repressor 2 are bound to the DNA:

$$F_{A_1A_2R_2} = \frac{C_{12}^{a} K_{A_1}[A_1] K_{A_2}[A_2] K_{R_2}[R_2]}{Z}$$

Activator 1, repressor 1 and 2 are bound to the DNA:

$$F_{A_1R_1R_2} = \frac{K_{A_1}[A_1] C_{12}^{r} K_{R_1}[R_1] K_{R_2}[R_2]}{Z}$$

Activator 2, repressor 1 and 2 are bound to the DNA:

$$F_{A_2R_1R_2} = \frac{K_{A_2}[A_2] C_{12}^{r} K_{R_1}[R_1] K_{R_2}[R_2]}{Z}$$

Activator 1 and 2 and repressor 1 and 2 are bound to the DNA:

$$F_{A_1A_2R_1R_2} = \frac{C_{12}^{a} K_{A_1}[A_1] K_{A_2}[A_2] C_{12}^{r} K_{R_1}[R_1] K_{R_2}[R_2]}{Z}$$

Then the states vector of two activators and two repressors can be written in the following way.

$$F = [F_N, F_{A_1}, F_{A_2}, F_{R_1}, F_{R_2}, F_{A_1A_2}, F_{A_1R_1}, F_{A_1R_2}, F_{A_2R_1}, F_{A_2R_2}, F_{R_1R_2},$$
$$F_{A_1A_2R_1}, F_{A_1A_2R_2}, F_{A_1R_1R_2}, F_{A_2R_1R_2}, F_{A_1A_2R_1R_2}],$$

In the expressions above we enumerate the Boltzmann states of an enhancer with two activator and two repressor binding sites. We claim that the binding of the repressors modulates the probability of states with activators and repressors simultaneously bound by a quenching factor. For example, binding of the activator A_j and repressor R_i simultaneously is reduced by a factor of $(1-q_{R_i}^{A_j})$. We can modify the probability of states, in the following way:

$$\begin{aligned}\tilde{F} = [&F_N, F_{A_1}, F_{A_2}, F_{R_1}, F_{R_2}, F_{A_1A_2}, F_{A_1R_1}(1-q_{R_1}^{A_1}), F_{A_1R_2}(1-q_{R_2}^{A_1}), F_{A_2R_1}(1-q_{R_1}^{A_2}),\\ &F_{A_2R_2}(1-q_{R_2}^{A_2}), F_{R_1R_2}, F_{A_1A_2R_1}(1-q_{R_1}^{A_1})(1-q_{R_1}^{A_2}), F_{A_1A_2R_2}(1-q_{R_2}^{A_1})(1-q_{R_2}^{A_2}),\\ &F_{A_1R_1R_2}(1-q_{R_1}^{A_1})(1-q_{R_2}^{A_1}), F_{A_2R_1R_2}(1-q_{R_1}^{A_2})(1-q_{R_2}^{A_2}),\\ &F_{A_1A_2R_1R_2}(1-q_{R_1}^{A_1})(1-q_{R_2}^{A_1})(1-q_{R_1}^{A_2})(1-q_{R_2}^{A_2})]\end{aligned}$$

With a description of states in hand, we calculate the total efficiency of the enhancer. Let E_{A_1} and E_{A_2} be efficiency of activators when they are bound to the DNA alone. We can write the case of two activators bound to the DNA at the same time in the following way. Here w_1^a and w_2^a are the weight constants.

$$E_{A_1A_2} = w_1^a E_{A_1} + w_2^a E_{A_2}$$

Let E_{R_1} and E_{R_2} be efficiency of activators when they are bound to the DNA alone. We can write the case of two repressors bound to the DNA at the same time in the following way. Here w_1^r and w_2^r are the weight constants.

$$E_{R_1R_2} = w_1^r E_{R_1} + w_2^r E_{R_2}$$

Then the overall efficiency vector of two activators and two repressors can be written in the following way:

$$\begin{aligned}
E = [&E_N, E_{A_1}, E_{A_2}, E_{R_1}, E_{R_2}, E_{A_1A_2}, E_{A_1R_1}, E_{A_1R_2}, E_{A_2R_1}, E_{A_2R_2}, E_{R_1R_2}, \\
&E_{A_1A_2R_1}, E_{A_1A_2R_2}, E_{A_1R_1R_2}, E_{A_2R_1R_2}, E_{A_1A_2R_1R_2}] \\
= [&E_N, E_{A_1}, E_{A_2}, E_{R_1}, E_{R_2}, w_1^a E_{A_1} + w_2^a E_{A_2}, E_{A_1} + E_{R_1}, E_{A_1} + E_{R_2}, \\
&E_{A_2} + E_{R_1}, E_{A_2} + E_{R_2}, w_1^r E_{R_1} + w_2^r E_{R_2}, w_1^a E_{A_1} + w_2^a E_{A_2} + E_{R_1}, \\
&w_1^a E_{A_1} + w_2^a E_{A_2} + E_{R_2}, E_{A_1} + w_1^r E_{R_1} + w_2^r E_{R_2}, E_{A_2} + w_1^r E_{R_1} + w_2^r E_{R_2}, \\
&w_1^a E_{A_1} + w_2^a E_{A_2} + w_1^r E_{R_1} + w_2^r E_{R_2}]
\end{aligned}$$

Then we can obtain the total expression level of the enhancer as the sum of the expression contribution of each state.

$Ex = \sum_i \tilde{F}_i T(E_i)$, where T is a regulatory function used to calculate expression contribution of each state. If we set the following simple assumptions $E_N = 0$, $E_{A_1} = 10$,

$E_{A_2} = 10$, $w_1^a = 0.5$, $w_2^a = 0.5$, $E_{R_1} = 0$, $E_{R_2} = 0$, and $T(x) = \frac{1}{1+e^{5-x}}$, the general framework described above reduces to a form employed by Zinzen *et al* (2006) and total expression of the enhancer with 2 activator and 2 repressor binding sites can be written as:

$$Ex \approx F_{A_1} + F_{A_2} + F_{A_1A_2} + F_{A_1R_1}(1-q_{R_1}^{A_1}) + F_{A_1R_2}(1-q_{R_2}^{A_1}) + F_{A_2R_1}(1-q_{R_1}^{A_2}) +$$
$$F_{A_2R_2}(1-q_{R_2}^{A_2}) + F_{A_1A_2R_1}(1-q_{R_1}^{A_1})(1-q_{R_1}^{A_2}) + F_{A_1A_2R_2}(1-q_{R_2}^{A_1})(1-q_{R_2}^{A_2}) +$$
$$F_{A_1R_1R_2}(1-q_{R_1}^{A_1})(1-q_{R_2}^{A_1}) + F_{A_2R_1R_2}(1-q_{R_1}^{A_2})(1-q_{R_2}^{A_2}) +$$
$$F_{A_1A_2R_1R_2}(1-q_{R_1}^{A_1})(1-q_{R_2}^{A_1})(1-q_{R_1}^{A_2})(1-q_{R_2}^{A_2})$$

Cooperative Binding:

Incorporating cooperative binding for the general case is not a simple matter to solve, since we might have all different cooperativity parameters for different protein types. We can have simple assumptions like all neighboring activators (repressors) are cooperating or all binding activators (repressors) are cooperating. The formulations of these schemes for homotypic cooperativity are explained in Zinzen & Papatsenko (2007).

We can also assume that cooperative binding level changes by distance such as proteins cooperate better if they are closer. This could be done by multiplying the cooperativity parameter C by a distance dependent function as described in Segal *et al* (2008), which used a Gaussian function with mean 0 and variance 50.

Combining fractional occupancy to expression:

The modeling of an enhancer and its fractional occupancy must also consider how those states ultimately affect transcription levels. If we assume that occupancy of the enhancer by the activators is the only indication for transcription and there is no need to specify different weights to different states we could use a simple linear function $T(x) = x$ (Zinzen & Papatsenko, 2007). However, numerous experimental studies have demonstrated nonlinearity in gene activation, which probably reflects the thresholds imposed by chromatin and inherent limits on transcriptional initiation rates. Thus a nonlinear function of sigmoidal type $T(x) = \frac{1}{1+e^{b-x}}$ can be used, as described in Segal *et al* (2008), where b represents basal expression level. We used $T(x) = \frac{1}{1+e^{5-x}}$ for our analysis on synthetic enhancers and *rhomboid* NEE.

Binding affinity:

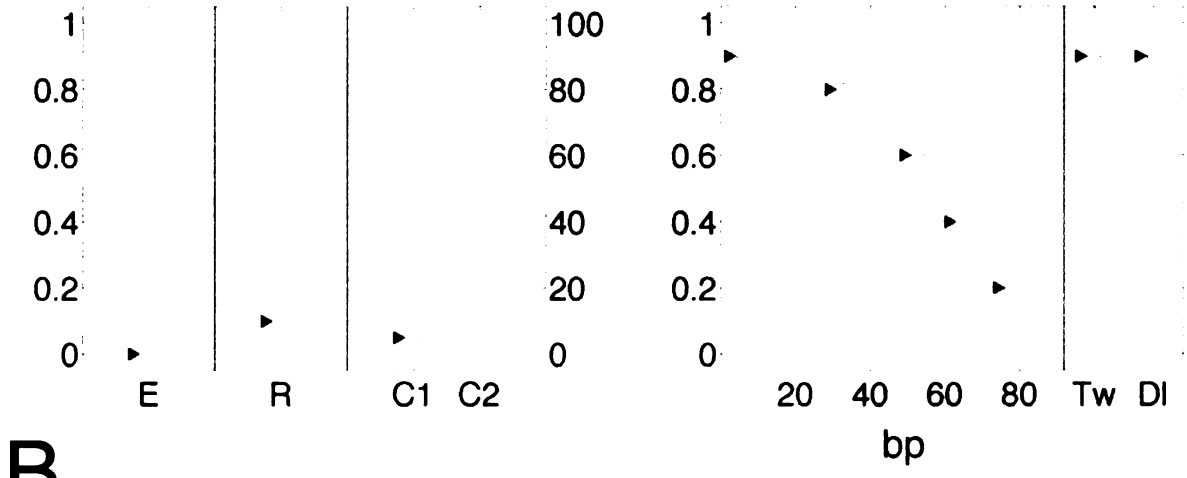
The set of synthetic gene modules used in this study did not explore the effects of different qualities of binding sites to any extent, but for analysis of endogenous regulatory regions, this consideration is paramount. In the *rhomboid* NEE study, for each transcription factor, we took the score of the strongest site among all those that bind that transcription factor as a free parameter and constrain the other values by treating the PWM score as a free energy of binding (Stormo, 2000).

Parameter Estimation:

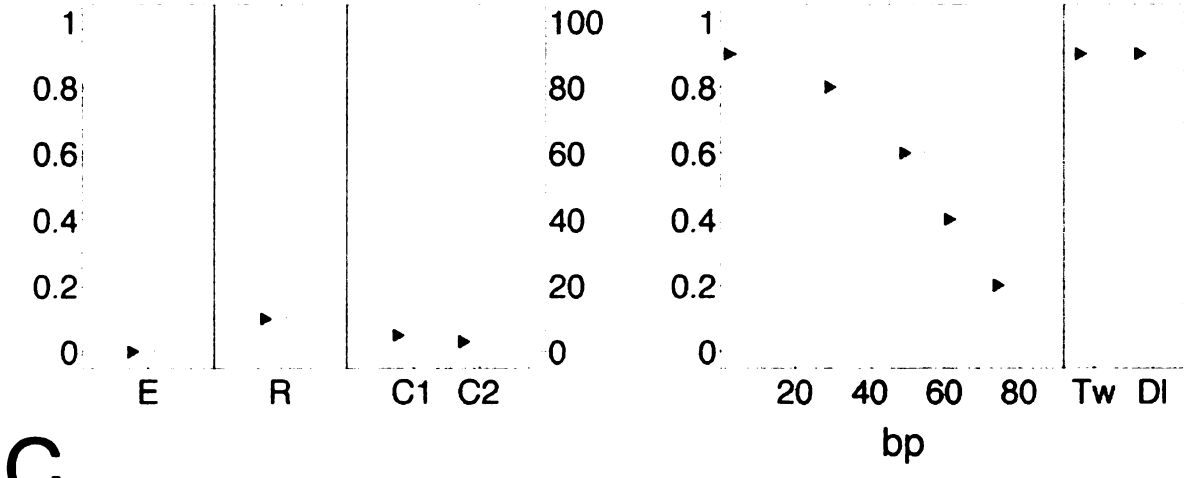
To fit the model with the actual data we minimized the root mean square error between the predicted and experimental data. Root mean square error is defined in the following way: $RMSE = (\frac{1}{n} \sum_{j=1}^{n} (y_p(j) - y_e(j))^2)^{1/2}$, where n is the number of data points, y_p is the model's prediction and y_e is the experimental data (e.g. Figure 3). The fitting is done by optimizing the parameters that gives the minimum RMSE error. Global parameter estimation algorithm, evolutionary strategy (ES) as described in Runarsson & Yao (2005) is used for estimation. ES is a nature-inspired parameter estimation method belonging to the class of evolutionary algorithms which uses mutation, and selection applied to a population of individuals containing candidate solutions in order to evolve iteratively better and better solutions. Three global parameter estimation techniques (genetic algorithm (GA), simulated annealing (SA) and evolutionary strategy) were checked for our model on synthetic data before we selected ES. The results of ES, GA and SA are shown in the following figures (Figure A-1-3) on synthetic data. The estimation is done for 9 different schemes and 100 runs and results are plotted. The ES algorithm clearly functioned better than genetic algorithms and simulated annealing on our model. For further testing of our parameter estimation algorithm, we added different amounts of noise to the synthetic data (1% to 30%). Adding noise to the synthetic data does not change the estimation results drastically (Figure A-4).

Figure A-1: One hundred parameter estimation runs are shown using the evolutionary strategy parameter estimation method for all schemes (variant formulations of the model employing different assumptions for cooperativity and quenching).

A


1
0.8
0.6
0.4
0.2
0
E
R
C1
C2
100
80
60
40
20
0
1
0.8
0.6
0.4
0.2
0
20
40
60
80
bp
Tw
Dl

B


1
0.8
0.6
0.4
0.2
0
E
R
C1
C2
100
80
60
40
20
0
1
0.8
0.6
0.4
0.2
0
20
40
60
80
bp
Tw
Dl

C

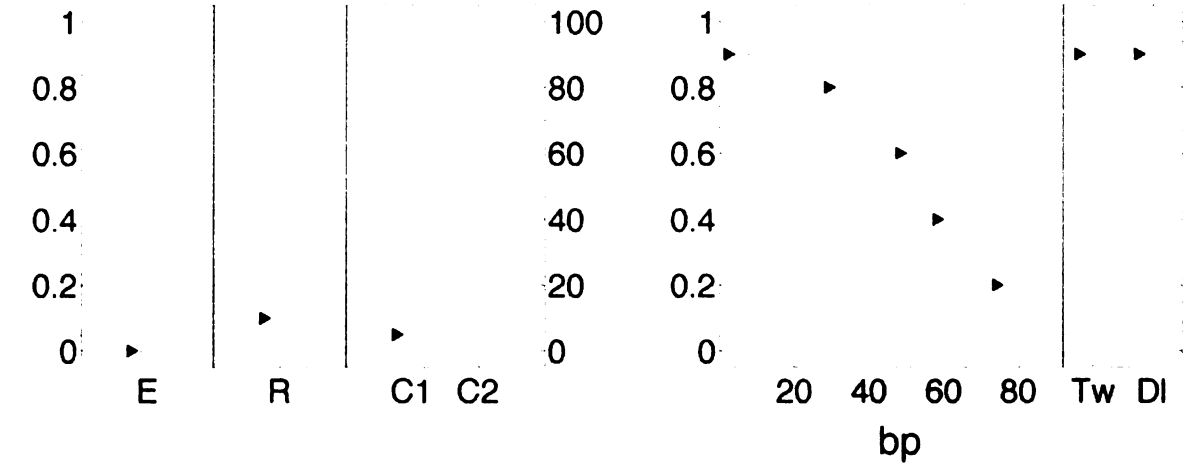
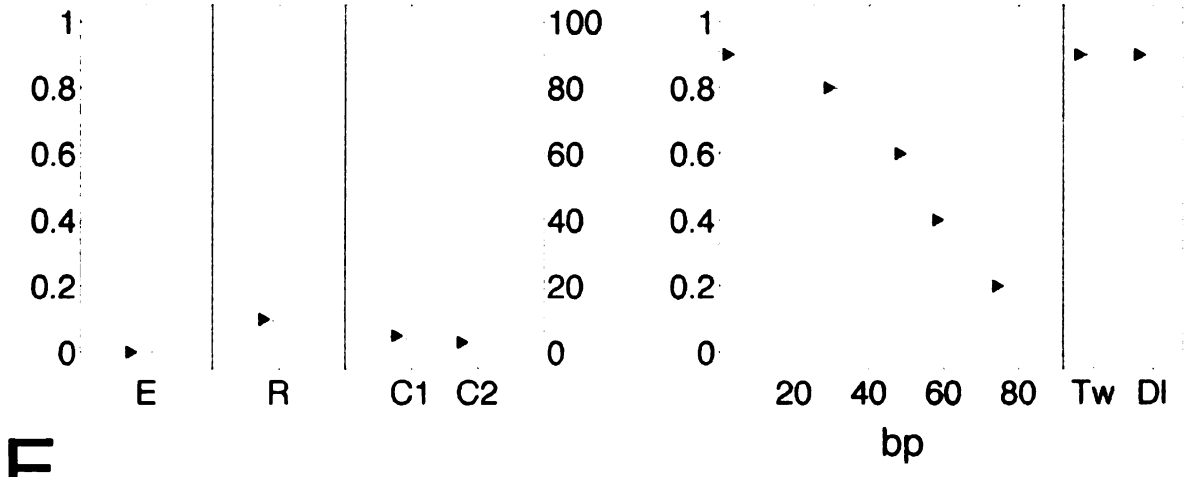
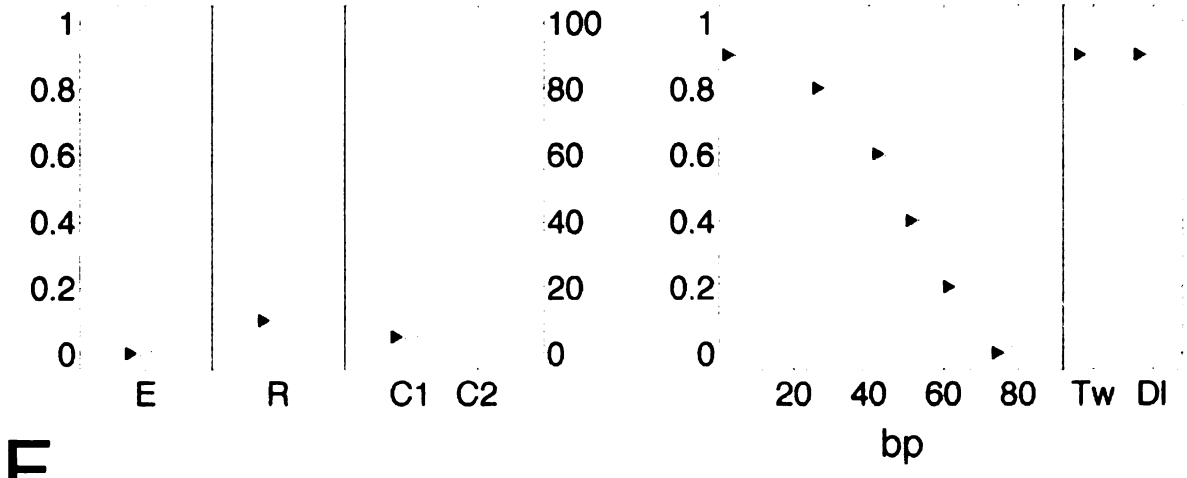

1
0.8
0.6
0.4
0.2
0
E
R
C1
C2
100
80
60
40
20
0
1
0.8
0.6
0.4
0.2
0
20
40
60
80
bp
Tw
Dl

Figure A-1: Continued.

D



E



F

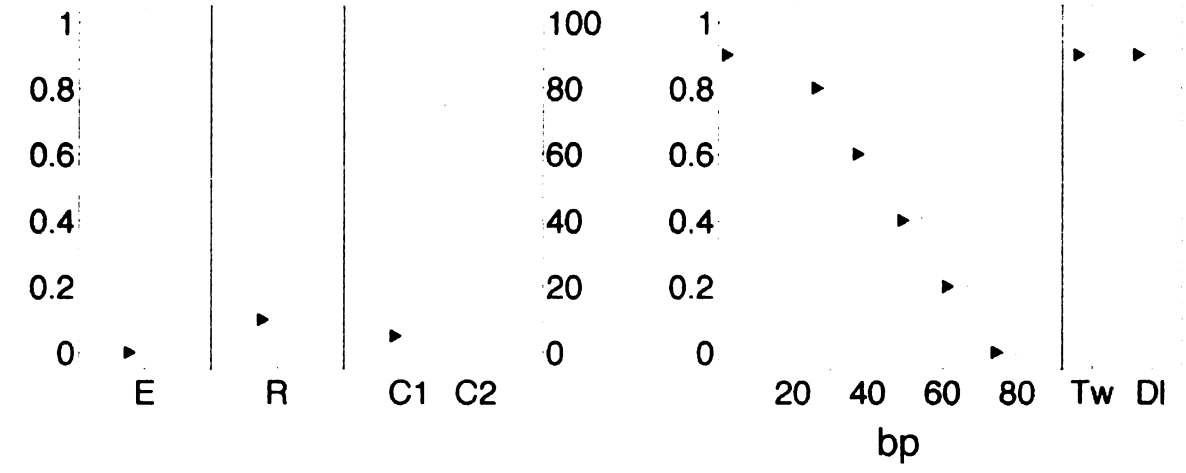
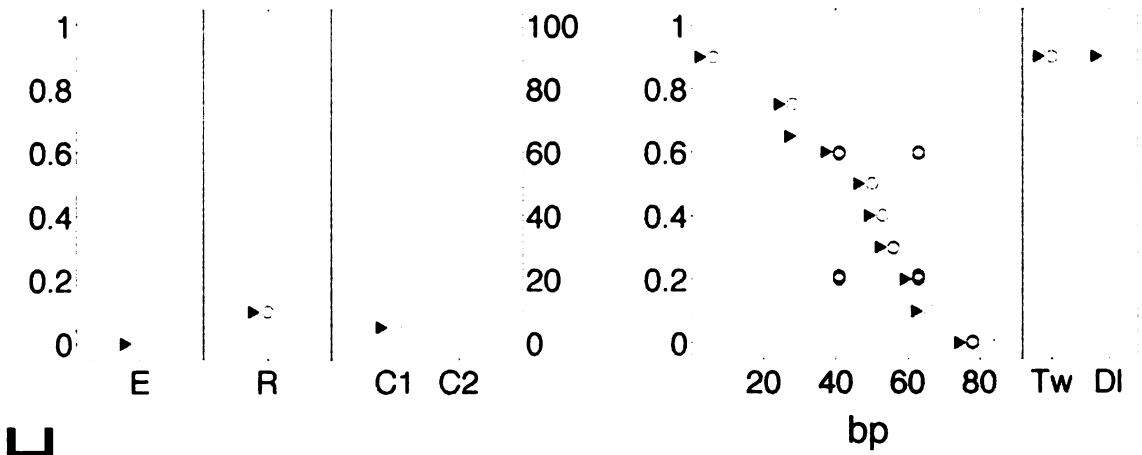
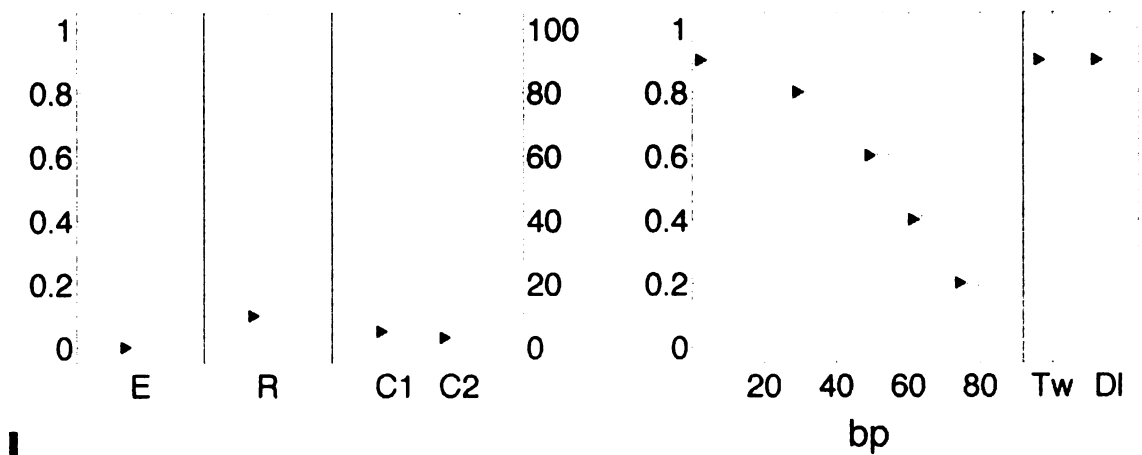


Figure A-1: Continued.

G



H



I

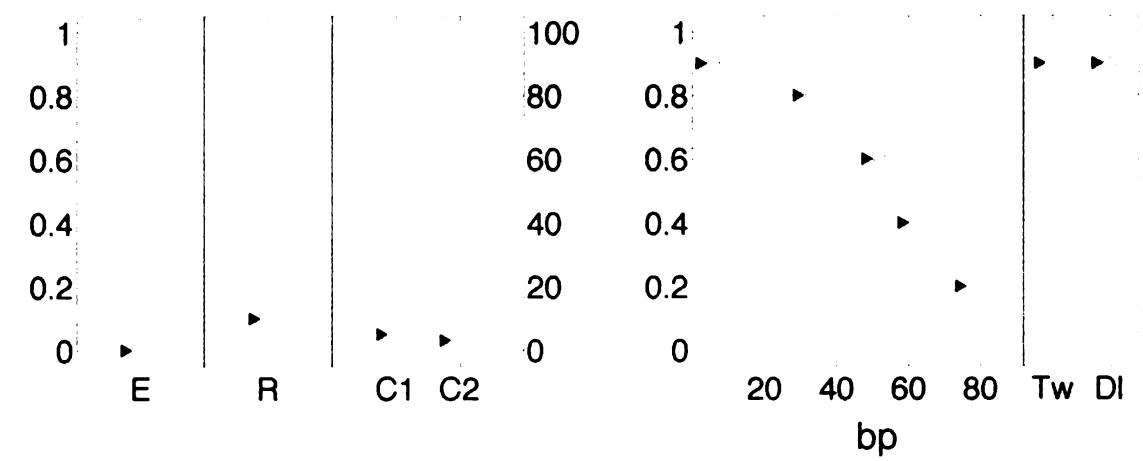


Figure A-2: One hundred parameter estimation runs are shown using the genetic algorithm method for all schemes.

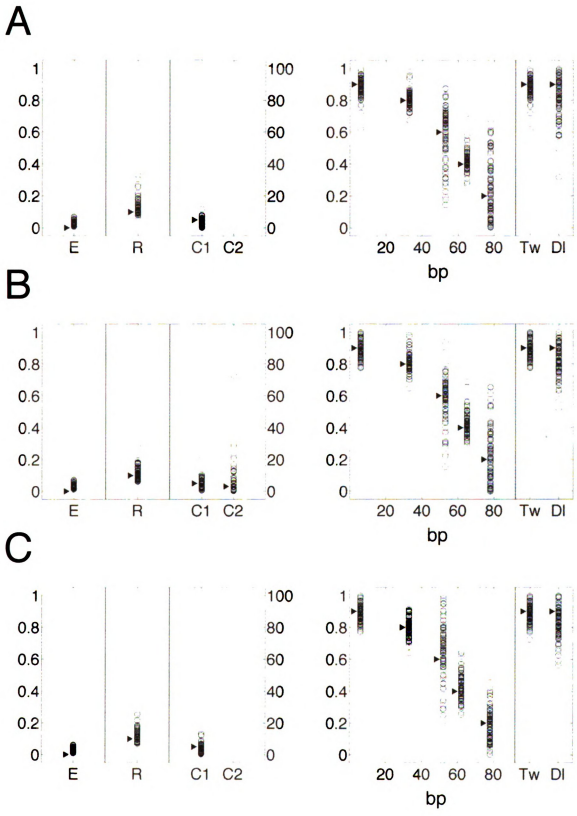
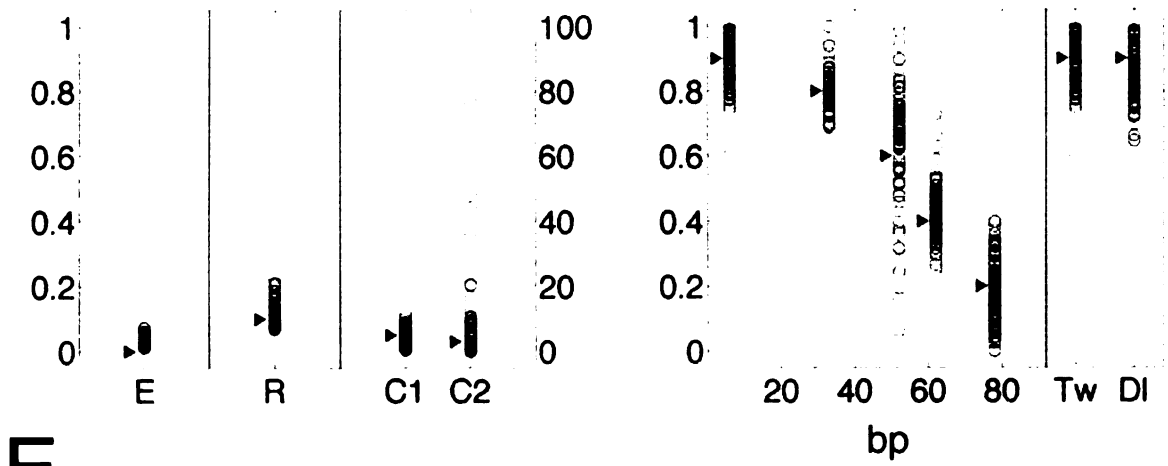
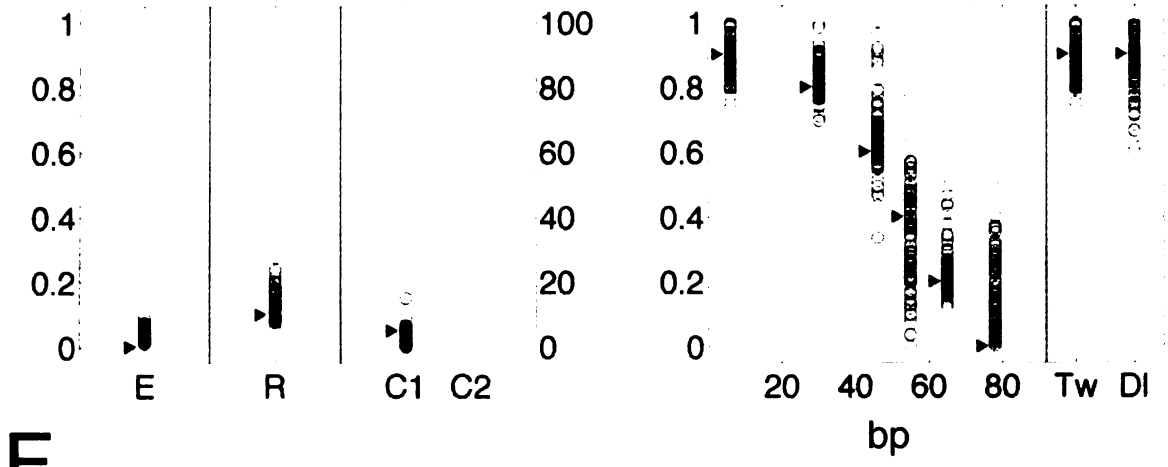

A
B
C
1
0.8
0.6
0.4
0.2
0
100
80
60
40
20
0
E
R
C1
C2
20
40
60
80
bp
Tw
Dl

Figure A-2: Continued.

D



E



F

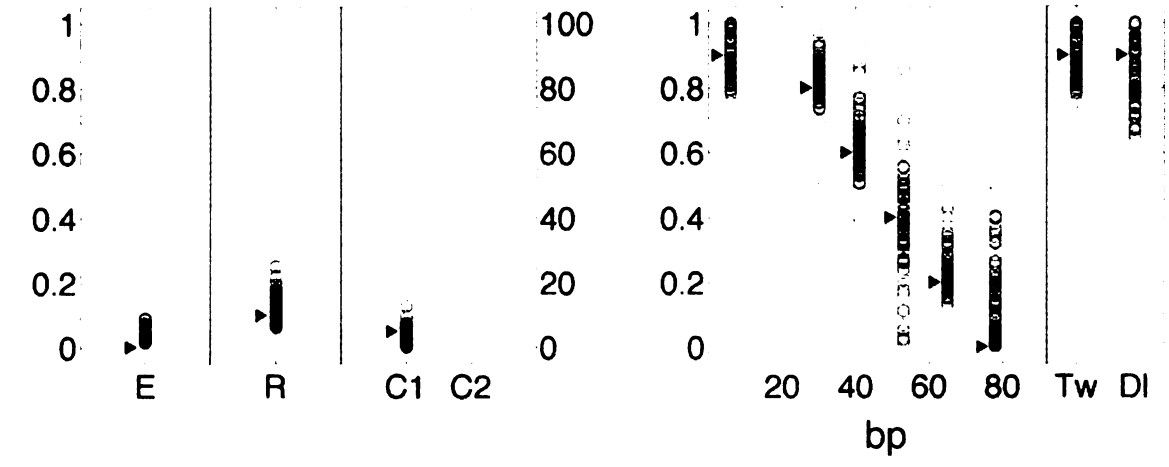
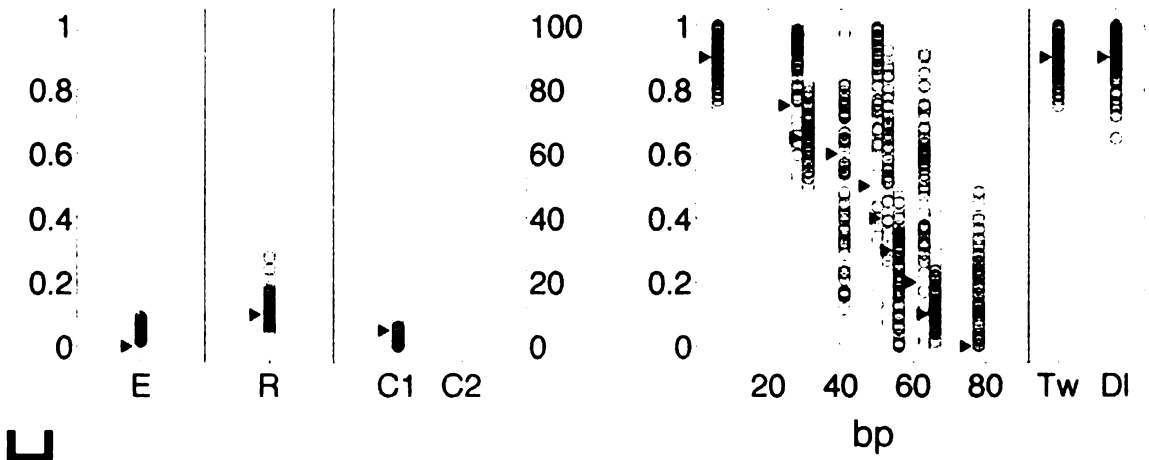
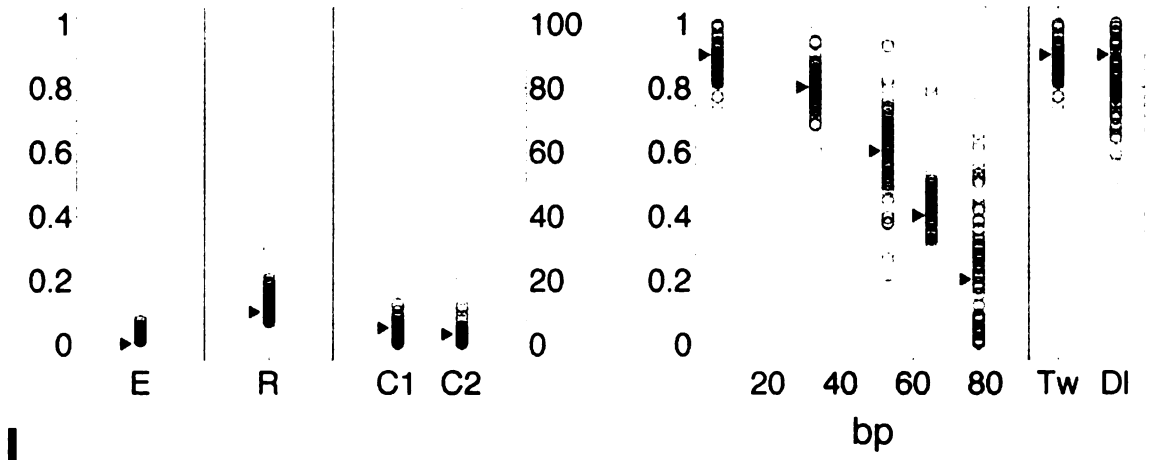


Figure A-2: Continued.

G



H



I

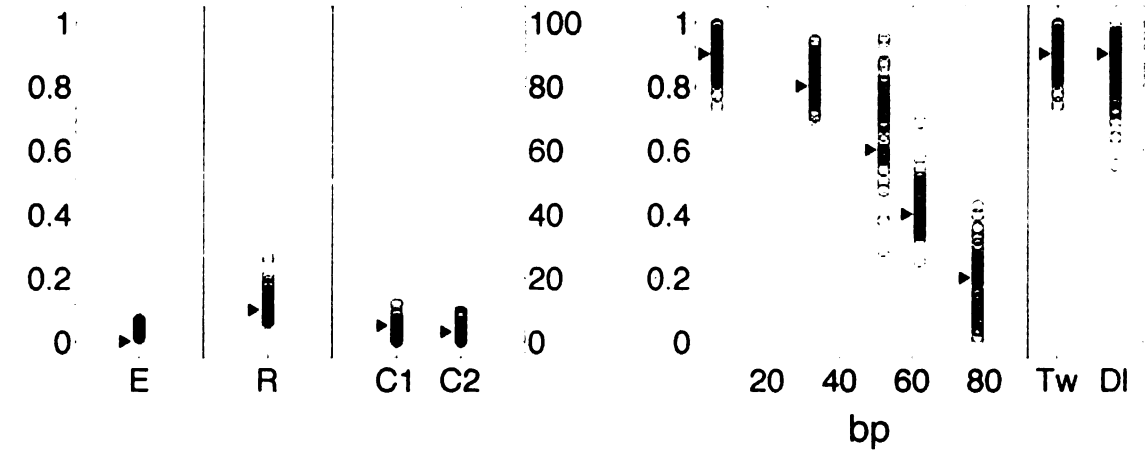


Figure A-3: One hundred parameter estimation runs are shown using the simulated annealing method for all schemes.

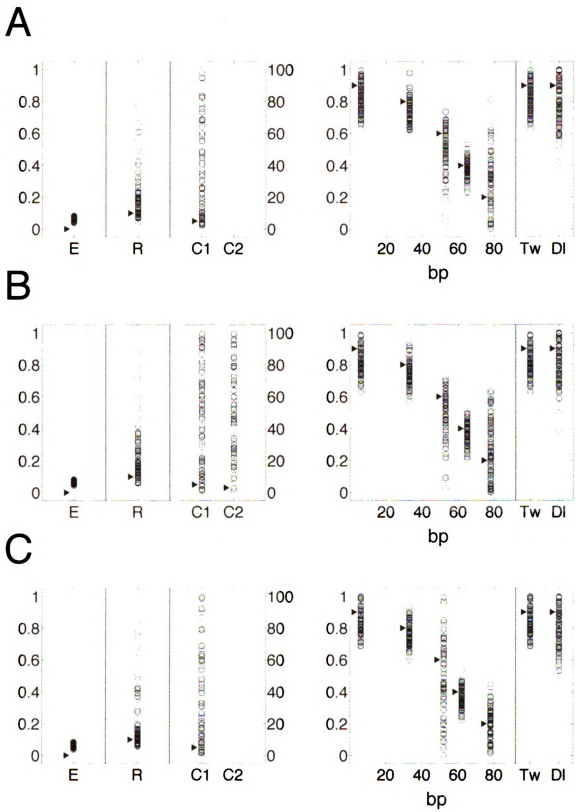

A
B
C
1
0.8
0.6
0.4
0.2
0
100
80
60
40
20
0
E
R
C1
C2
20
40
60
80
bp
Tw
DI

Figure A-3: Continued.

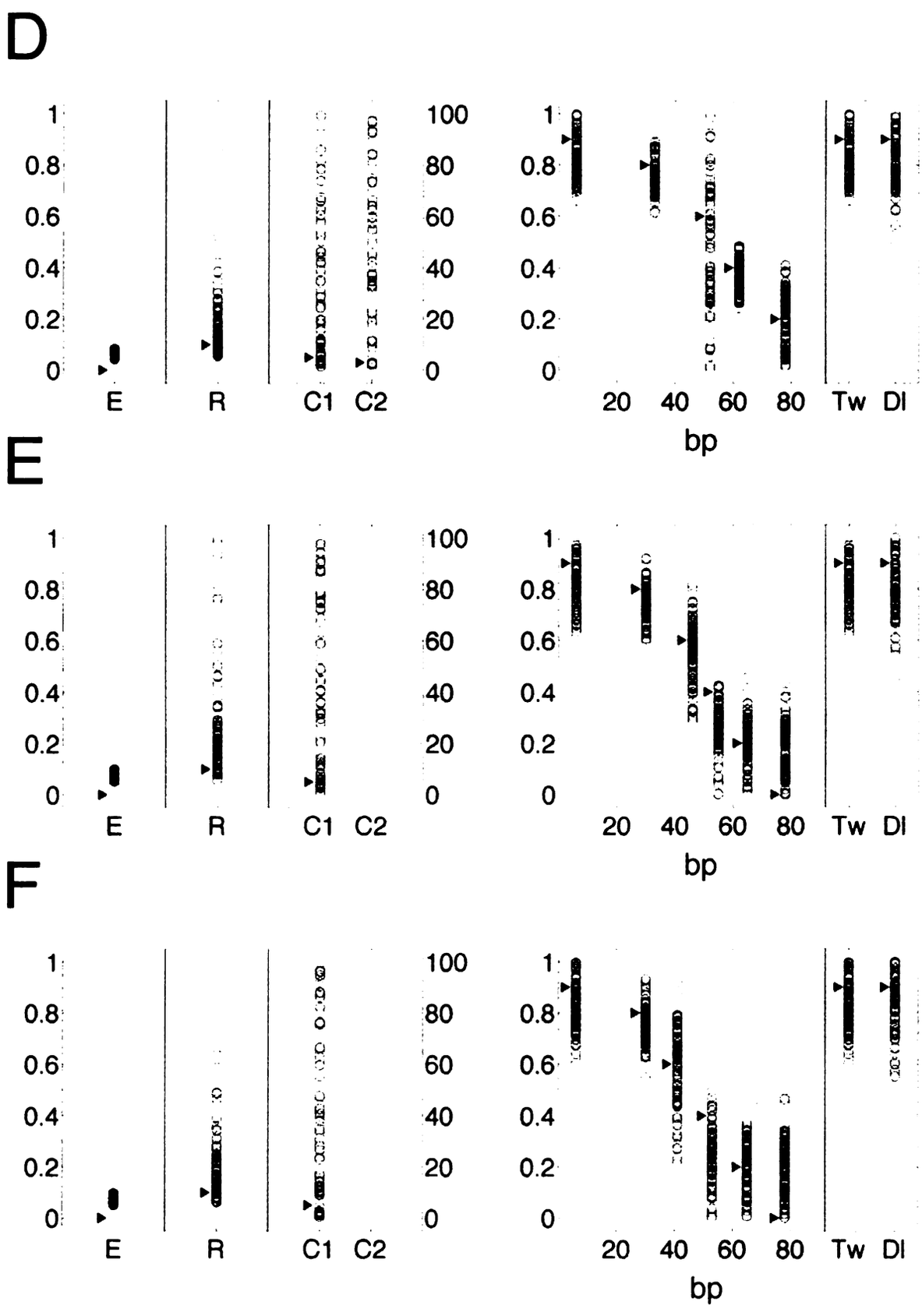
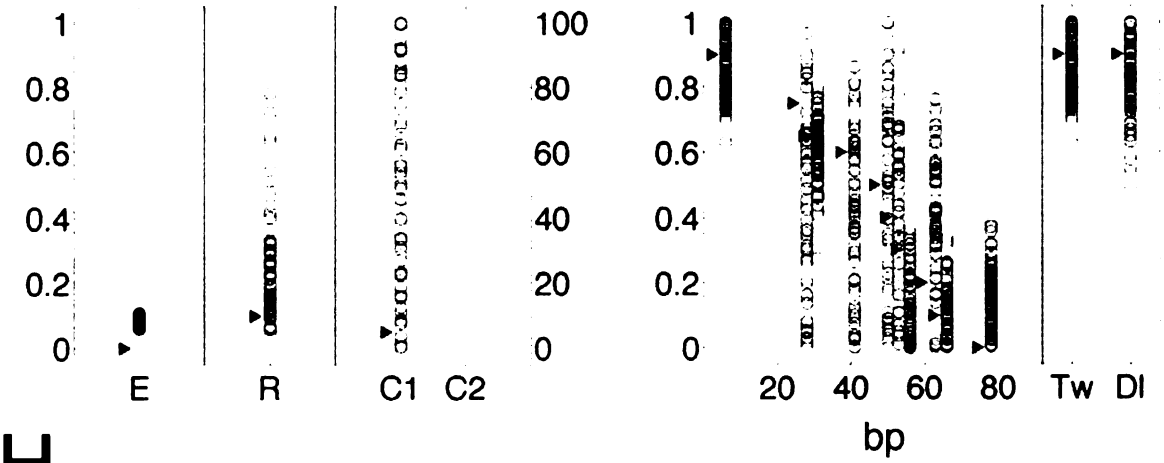
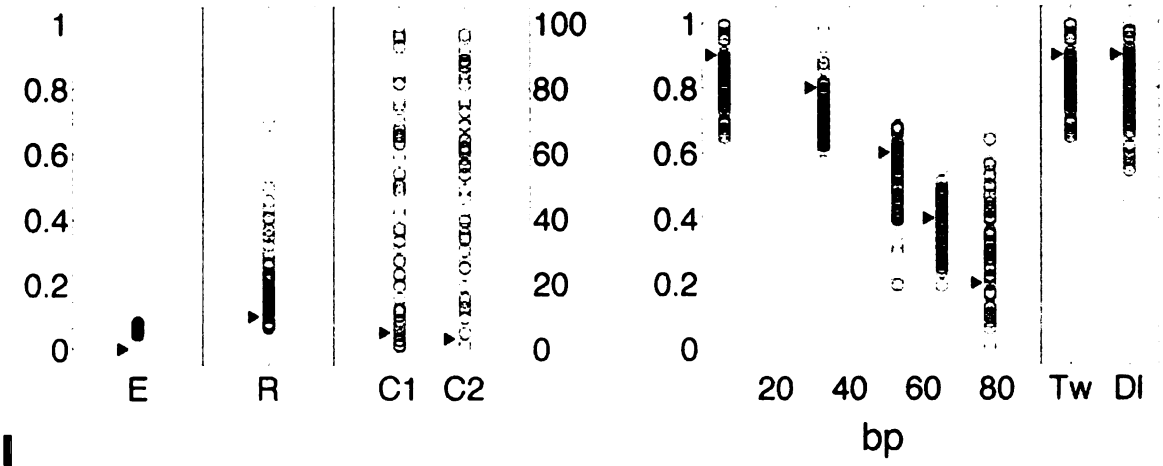


Figure A-3: Continued.

G



H



I

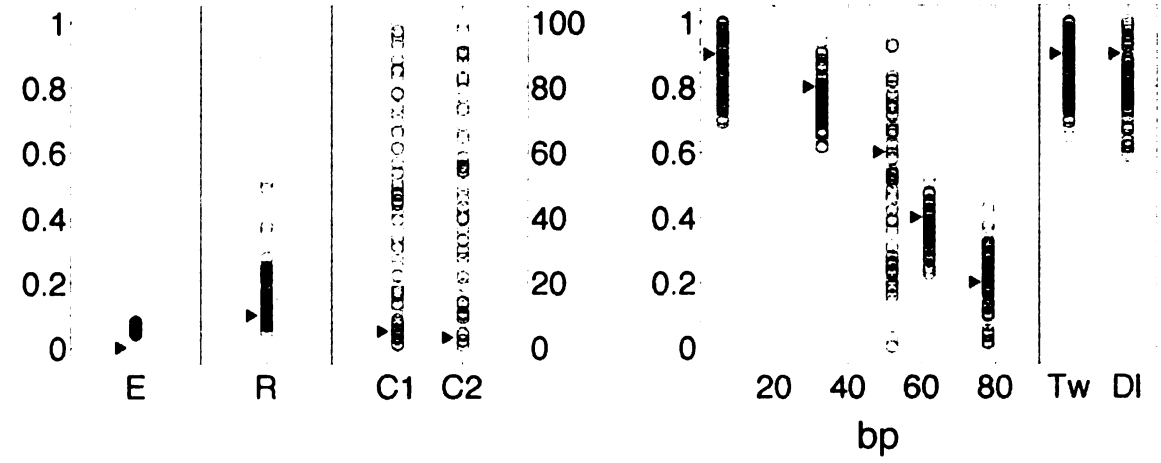
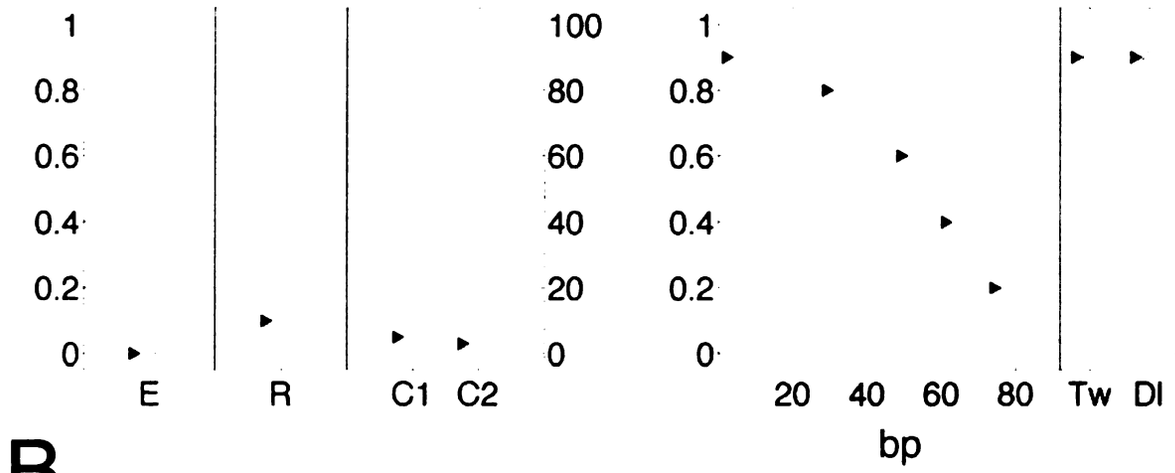
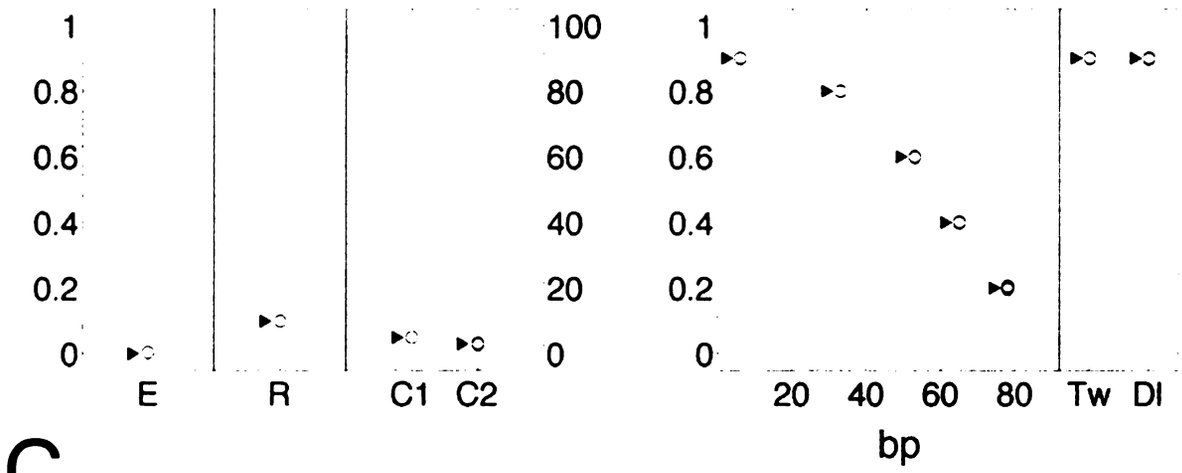


Figure A-4: Robustness of the evolutionary strategy parameter estimation technique on synthetic data with added noise. (A) No noise, (B) 1% noise, (C) 3% noise, (D) 5% noise, (E) 10% noise, (F) 20% noise and (G) 30% noise.

A


1
0.8
0.6
0.4
0.2
0
E
R
C1
C2
100
80
60
40
20
0
1
0.8
0.6
0.4
0.2
0
20
40
60
80
bp
Tw
DI

B


1
0.8
0.6
0.4
0.2
0
E
R
C1
C2
100
80
60
40
20
0
1
0.8
0.6
0.4
0.2
0
20
40
60
80
bp
Tw
DI

C

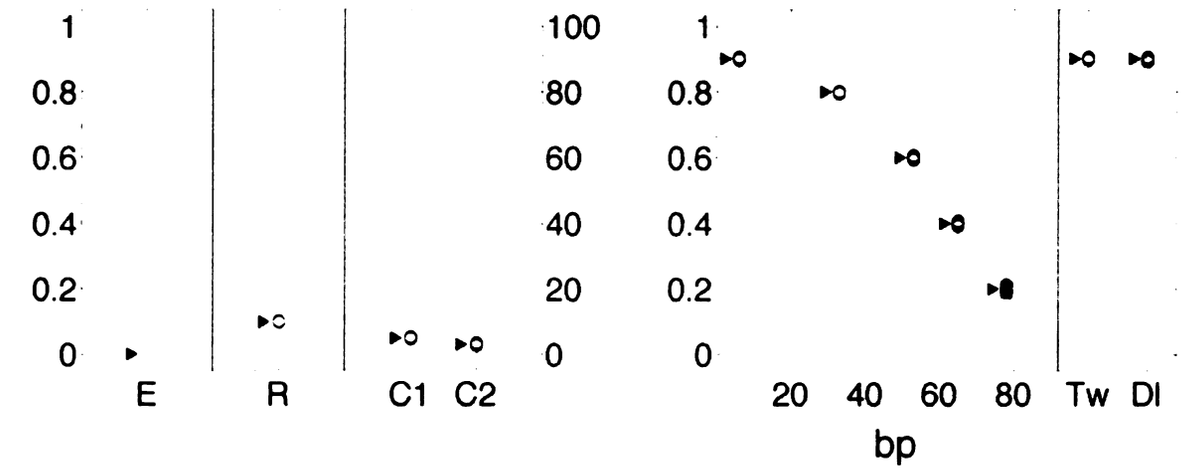
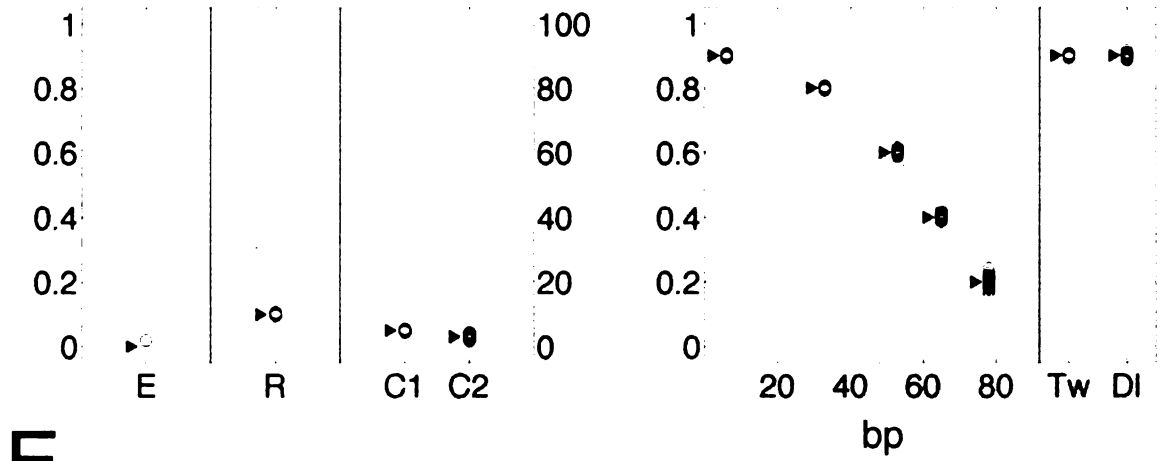
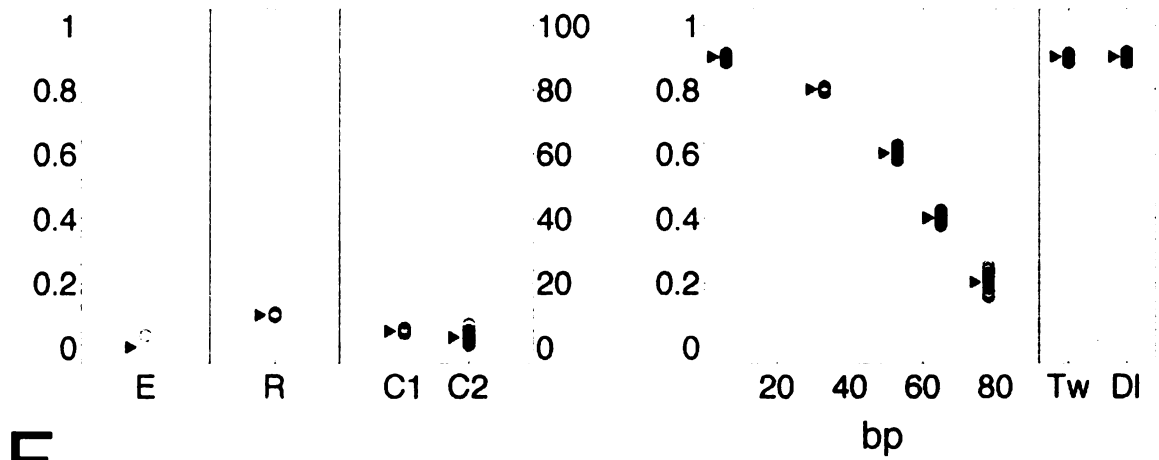

1
0.8
0.6
0.4
0.2
0
E
R
C1
C2
100
80
60
40
20
0
1
0.8
0.6
0.4
0.2
0
20
40
60
80
bp
Tw
DI

Figure A-4: Continued.

D



E



F

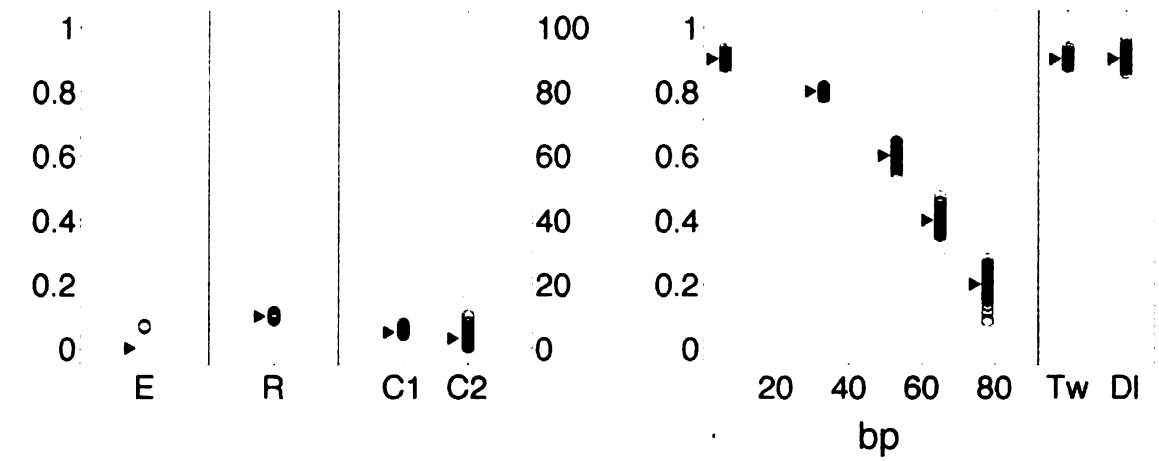


Figure A-4: Continued.

G

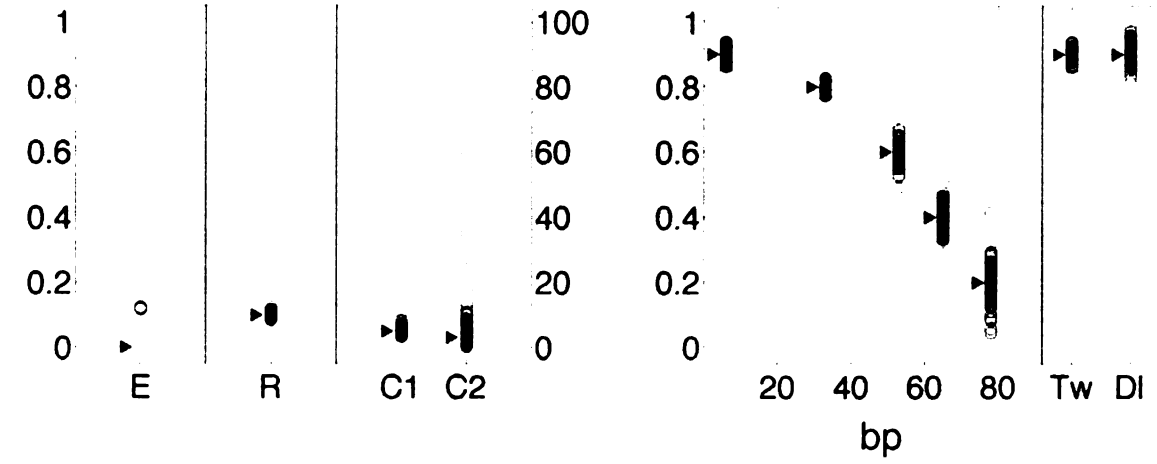


Figure A-5: Results of parameter estimation for R, C_1, C_2 and quenching employing different schemes. (A) scheme 1, (B) scheme 2, (C) scheme 3, (D) scheme 4, (E) scheme 5, (F) scheme 6, (G) scheme 7, (H) scheme 8 and (I) scheme 9.

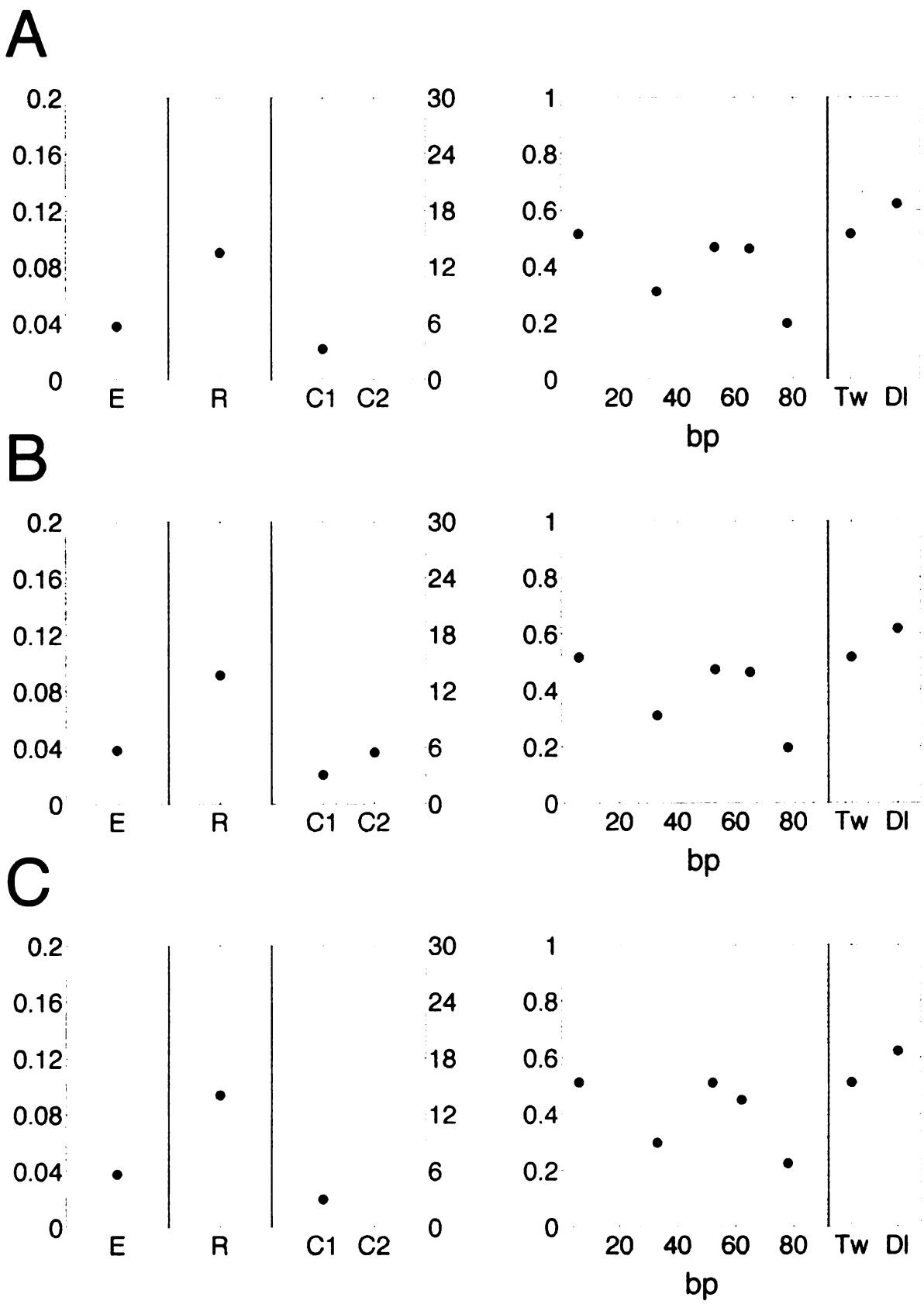

A
0.2
0.16
0.12
0.08
0.04
0
E
R
C1
C2
30
24
18
12
6
0
1
0.8
0.6
0.4
0.2
0
20
40
60
80
bp
Tw
Dl
B
0.2
0.16
0.12
0.08
0.04
0
E
R
C1
C2
30
24
18
12
6
0
1
0.8
0.6
0.4
0.2
0
20
40
60
80
bp
Tw
Dl
C
0.2
0.16
0.12
0.08
0.04
0
E
R
C1
C2
30
24
18
12
6
0
1
0.8
0.6
0.4
0.2
0
20
40
60
80
bp
Tw
Dl

Figure A-5: Continued.

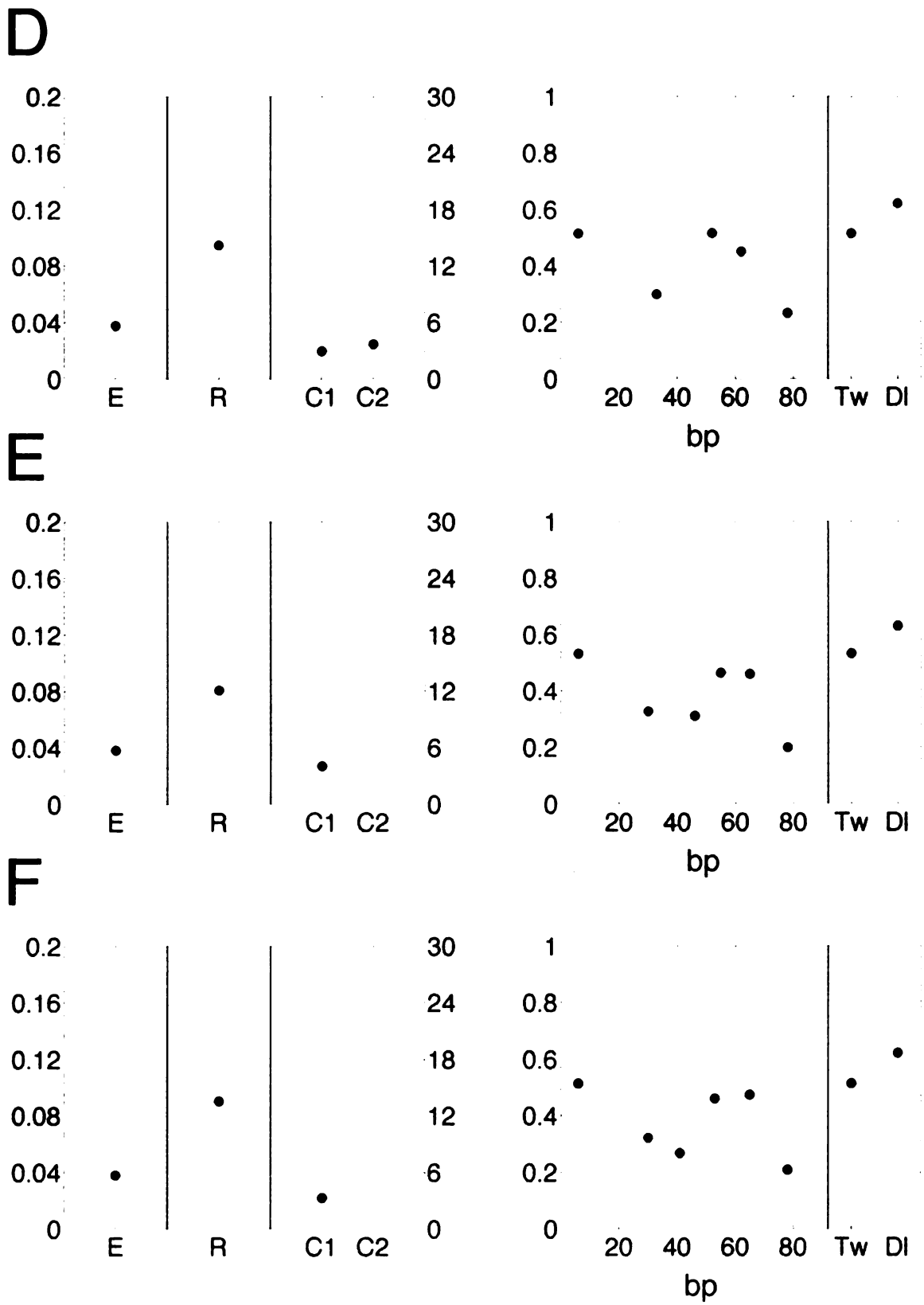


Figure A-5: Continued.

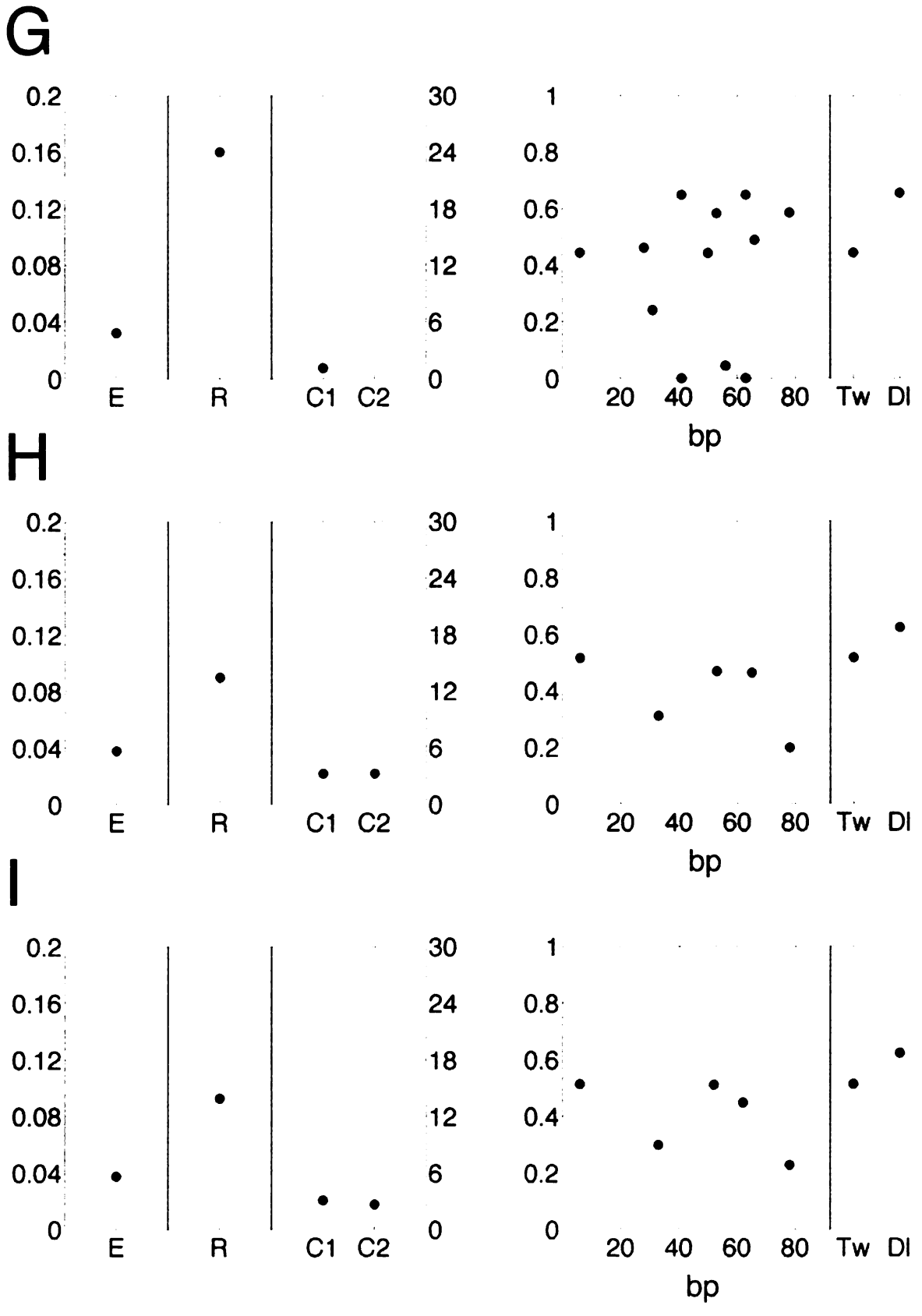


Figure A-6: Comparison of average error induced by leave-one-out analysis, employing nine different formulations of the model. Lower error levels were noted with schemes 1-4, 8 and 9.

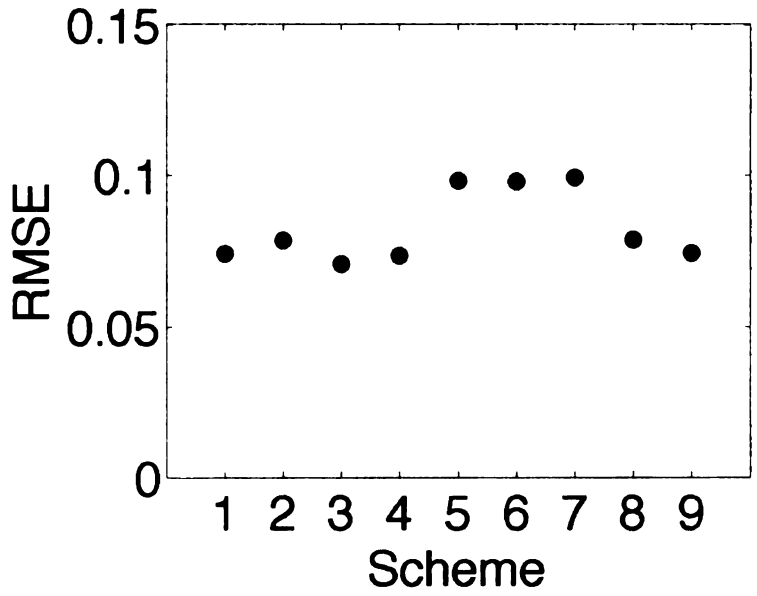


Table A-I: Ex functions for the synthetic enhancers of the model.

2gt.2t2d 2gt.25.2t2d 2gt.35.2t2d 2gt.50.2t2d 2gt.60.2t2d	$$Ex = \frac{S_A[A]}{1+S_A[A]} \times \frac{1+(2-q_1-q_2)S_R[Gt]+C_{12}^{Gt}(1-q_1)(1-q_2)(S_R[Gt])^2}{1+2S_R[Gt]+C_{12}^{Gt}(S_R[Gt])^2}$$ For Scheme 1-9 $C_{12}^{Gt} = C_1$
1gt.2t2d 1gt.25.2t2d	$$Ex = \frac{S_A[A]}{1+S_A[A]} \times \frac{1+(1-q_1)S_R[Gt]}{1+S_R[Gt]}$$
3gt.2t2d 3gt.50.2t2d	$$Ex = \frac{S_A[A]}{1+S_A[A]} \times \frac{\{1+(3-q_1-q_2-q_3)S_R[Gt]+(C_{12}^{Gt}(1-q_1)(1-q_2)+C_{13}^{Gt}(1-q_1)(1-q_3)+ C_{23}^{Gt}(1-q_2)(1-q_3))(S_R[Gt])^2+C_{123}^{Gt}(1-q_1)(1-q_2)(1-q_3)(S_R[Gt])^3\}}{1+3S_R[Gt]+(C_{12}^{Gt}+C_{13}^{Gt}+C_{23}^{Gt})(S_R[Gt])^2+C_{123}^{Gt}(S_R[Gt])^3}$$ For Scheme 1, 3, 5, 6, and 7 $C_{12}^{Gt} = C_1,\ C_{13}^{Gt} = C_1,\ C_{23}^{Gt} = C_1$ and $C_{123}^{Gt} = C_1^2$ For Scheme 2 and 4 $C_{12}^{Gt} = C_1,\ C_{13}^{Gt} = C_2,\ C_{23}^{Gt} = C_1$ and $C_{123}^{Gt} = C_1 + C_2$ For Scheme 8 and 9 $C_{12}^{Gt} = C_1,\ C_{13}^{Gt} = C_2,\ C_{23}^{Gt} = C_1$ and $C_{123}^{Gt} = C_1 \cdot C_2$

Table A-I: Continued.

2gt.2t2d.1gt	$$Ex = \frac{S_A[A]}{1+S_A[A]} \times \frac{\{1+(3-q_1-q_2-q_3)S_R[Gt]+(C_{12}^{Gt}(1-q_1)(1-q_2)+(1-q_1)(1-q_3)+(1-q_2)(1-q_3))(S_R[Gt])^2+C_{12}^{Gt}(1-q_1)(1-q_2)(1-q_3)(S_R[Gt])^3\}}{1+3S_R[Gt]+(C_{12}^{Gt}+1+1)(S_R[Gt])^2+C_{12}^{Gt}(S_R[Gt])^3}$$ For Scheme 1-9 $C_{12}^{Gt} = C_1$
1gt.2t2d.1gt 1gt.50.2t2d.1gt	$$Ex = \frac{S_A[A]}{1+S_A[A]} \times \frac{1+(2-q_1-q_2)S_R[Gt]+((1-q_1)(1-q_2))(S_R[Gt])^2}{1+2S_R[Gt]+(S_R[Gt])^2}$$

Table A-II: Parameter assignments for all the genes and schemes. Capital qs represent the quenching efficiency of repressors shown in the gene structure and lower case qs represent the assigned parameters, which are grouped into different 'bins' of distances as explained in Materials and Methods.

Gene	S1	S2	S3	S4	S5	S6	S7	S8	S9	Gene Structure
1	Q1=q1 Q2=q2	Q1=q1 Q2=q2	Q1=q1 Q2=q2	Q1=q1 Q2=q2	Q1=q1 Q2=q2	Q1=q1 Q2=q2	Q1=q1 Q2=q2	Q1=q1 Q2=q2	Q1=q1 Q2=q2	
2	Q1=q2 Q2=q3	Q1=q2 Q2=q3	Q1=q2 Q2=q3	Q1=q2 Q2=q3	Q1=q2 Q2=q4	Q1=q2 Q2=q4	Q1=q3 Q2=q6	Q1=q2 Q2=q3	Q1=q2 Q2=q3	
3	Q1=q2 Q2=q4	Q1=q2 Q2=q4	Q1=q2 Q2=q4	Q1=q2 Q2=q4	Q1=q3 Q2=q5	Q1=q3 Q2=q5	Q1=q4 Q2=q8	Q1=q2 Q2=q4	Q1=q2 Q2=q4	C1 Q1 Q2
4	Q1=q3 Q2=q5	Q1=q3 Q2=q5	Q1=q4 Q2=q5	Q1=q4 Q2=q5	Q1=q4 Q2=q6	Q1=q4 Q2=q6	Q1=q7 Q2=q10	Q1=q3 Q2=q5	Q1=q4 Q2=q5	
5	Q1=q4 Q2=0	Q1=q4 Q2=0	Q1=q4 Q2=0	Q1=q4 Q2=0	Q1=q5 Q2=0	Q1=q5 Q2=0	Q1=q9 Q2=0	Q1=q4 Q2=0	Q1=q4 Q2=0	
9	Q1=q1	Q1=q1	Q1=q1	Q1=q1	Q1=q1	Q1=q1	Q1=q1	Q1=q1	Q1=q1	
16	Q1=q2	Q1=q2	Q1=q2	Q1=q2	Q1=q2	Q1=q2	Q1=q3	Q1=q2	Q1=q2	Q1
10	Q1=q1 Q2=q2 Q3=q3	Q1=q1 Q2=q2 Q3=q3	Q1=q1 Q2=q2 Q3=q3	Q1=q1 Q2=q2 Q3=q3	Q1=q1 Q2=q2 Q3=q3	Q1=q1 Q2=q2 Q3=q4	Q1=q1 Q2=q2 Q3=q5	Q1=q1 Q2=q2 Q3=q3	Q1=q1 Q2=q2 Q3=q3	
17	Q1=q3 Q2=q5 Q3=0	Q1=q3 Q2=q5 Q3=0	Q1=q4 Q2=q5 Q3=0	Q1=q4 Q2=q5 Q3=0	Q1=q4 Q2=q6 Q3=0	Q1=q4 Q2=q6 Q3=0	Q1=q7 Q2=q10 Q3=0	Q1=q3 Q2=q5 Q3=0	Q1=q4 Q2=q5 Q3=0	C1 Q3 C1 Q1 C2 Q2
12	Q1=q1 Q2=q2 Q3=q6	Q1=q1 Q2=q2 Q3=q6	Q1=q1 Q2=q2 Q3=q6	Q1=q1 Q2=q2 Q3=q6	Q1=q1 Q2=q2 Q3=q7	Q1=q1 Q2=q2 Q3=q7	Q1=q1 Q2=q2 Q3=q11	Q1=q1 Q2=q2 Q3=q6	Q1=q1 Q2=q2 Q3=q6	C1 Q1 Q3 Q2
14	Q1=q1 Q2=q6	Q1=q1 Q2=q6	Q1=q1 Q2=q6	Q1=q1 Q2=q6	Q1=q1 Q2=q7	Q1=q1 Q2=q7	Q1=q1 Q2=q11	Q1=q1 Q2=q6	Q1=q1 Q2=q6	Q1 Q2
19	Q1=q3 Q2=q6	Q1=q3 Q2=q6	Q1=q4 Q2=q6	Q1=q4 Q2=q6	Q1=q4 Q2=q7	Q1=q4 Q2=q7	Q1=q7 Q2=q11	Q1=q3 Q2=q6	Q1=q4 Q2=q6	

Table A-III: Oligo and primer sequences used in this study.

1.	DA782/783 (2 Twist)	5'-GATCC**CATATG**TTGAG**CATATG**T 5'-CTAGACATATGCTCAACATATGG
2.	DA784/785 (2 Dorsal)	5'-CTAGA**GGGATTTTCCCA**AATCGA **GGGAAAACCCAA**CCGC 5'-GGTTGGGTTTTCCCTCGATTTGGGAAAATCC CT
3.	DA786/787 (25bp spacer)	5'-GATCTGGTTAGTAAGCTGTAAACTG 5'-GATCCAGTTTACAGCTTACTAACCA
4.	DA792/793 (2 Giant)	5'-AATTC**TATGACGCAAGA**ATGCGACTCG **TATGACGCAAGA**G 5'-GATCCTCTTGCGTCATACGAGTCGCATTCTTG CGTCATAG
5.	DA1255/1256 (1 Gt-EcoRI)	5'-AATTC**TATGACGCAAGA**ATGCGT 5'-AATTACGCATTCTTGCGTCATAG
6.	DA1257/1258 (1gt-EcoRI/BamHI)	5'-AATTC**TATGACGCAAGA**G 5'-GATCCTCTTGCGTCATAG
7.	DA(1259/1260) (1 gt-SacII)	5'-TG**TATGACGCAAGA**CCGC 5'-GGTCTTGCGTCATACAGA
8.	DA1403/1404 (2Dl.2Tw)	5'-AATTA**CATATG**TTGAG**CATATG**TCTAGA **TGGGAAAATCCC**TCGATT**TTGGGTTTTCCC**G 5'AATTCGGGAAAACCCAAAATCGAGGGATTTTCC CATCTAGACATATGCTCAACATATGT
9.	DA1405/1406 (50bp spacer)	5'-GATCTGGTTAGTAAGCTGTAAACTGGATCTGG TTAGTAAGCTGTAAACTG 5'-GATCCAGTTTACAGCTTACTAACCAGATCCAG TTTACAGCTTACTAACCA

Table A-III: Continued.

10.	DA1407/1408 (50bp spacer/1Gt)	5'-TGGTTAGTAAGCTGTAAACTGGATCTGGTTAG TAAGCTGTAAACTG**TATGACGCAAGA**CCGC 5'-GGTCTTGCGTCATACAGTTTACAGCTTACTAA CCAGATCCAGTTTACAGCTTACTAACCAGC
11.	DA1337/1338 (1 Krueppel)	5'-AATTC**AAAACGGGTTAAGC**G 5'-GATCCGCTTAACCCGTTTTG
12.	DA1339/1340 (2 Kr)	5'-AATTC**AAAACGGGTTAAGC**GACCC **AAAACGGGTTAAGC**G 5'-GATCCGCTTAACCCGTTTTGGGTCGCTTAACC CGTTTTG
13.	DA1341/1342 (3 Kr)	5'-AATTC**AAAACGGGTTAAGC**GACCC **AAAACGGGTTAAGC**GACCC**AAAACGGGTTAAG CG** 5'- GATCCGCTTAACCCGTTTTGGGTCGCTTAACCCG TTTTGGGTCGCTTAACCCGTTTTG
14.	DA1352/1353 (Kni.340bp.SpeI. Spacer)	5'-ACAT**ACTAGT**AACCGCTTTAGTCCCGCCAG 5'-ACATACTAGTTGTGCACGGAGCTCCGCGAG
15.	DA1833/1834 (20bp spacer.BamHI)	5'-GATCTGGTTAGTAAGCTGTG 5'-GATCCACAGCTTACTAACCA
16.	DA1835/1836 (35bp spacer.BamHI)	5'-GATCTGGTTAGTAAGCTGTAAACTCTGGTTAG TAG 5'-GATCCTACTAACCAGAGTTTACAGCTTACTAA CCA
17.	DA1837/1838 (Gt.(af).EcoRI)	5'-AATTCT**TGCGACGCATCA**ATGCGT 5'-AATTACGCATTGATGCGTCGCAAG
18.	DA1839/1840 (2Gt.(af).EcoRI/BamHI)	5'-AATTCT**TGCGACGCATCA**ATGCGACTCGT **TGCGACGCATCA**G 5'-GATCCTGATGCGTCGCAACGAGTCGCATTGAT GCGTCGCAAG

Table A-IV: Robustness of Evolutionary Strategy (ES) parameter estimation. The parameter estimation algorithms were run 100 times using the scheme 2 form of the model on a synthetic data set with increasing amounts of noise. Parameter value (Standard deviation). Results from parameter estimation using Genetic Algorithm (GA) and Simulated Annealing (SA) are shown for comparison.

	Exp.	ES (0%)	GA (0%)	SA (0%)	ES (1%)	ES (3%)	ES (5%)	ES (10%)	ES (20%)	ES (30%)
RMSE	0	0 (0)	0.033 (0.013)	0.066 (0.008)	0.004 (0)	0.012 (0)	0.019 (0)	0.037 (0)	0.067 (0.001)	0.119 (0.001)
R	10	10 (0)	11.853 (3.976)	22.006 (15.18)	9.999 (0.023)	9.999 (0.064)	9.997 (0.132)	9.996 (0.209)	9.981 (0.475)	10.038 (0.745)
C1	5	5 (0)	4.302 (3.328)	47.693 (30.536)	5.001 (0.028)	5.009 (0.11)	5.022 (0.165)	5.022 (0.326)	5.21 (0.663)	5.169 (1.109)
C2	3	3 (0)	6.334 (11.637)	50.118 (28.039)	3.006 (0.097)	3.03 (0.254)	3.07 (0.484)	3.418 (1.272)	4.11 (2.672)	5.81 (7.454)
Q1	0.9	0.9 (0)	0.897 (0.061)	.812 (0.086)	0.899 (0.001)	0.899 (0.001)	0.900 (0.002)	0.899 (0.006)	0.902 (0.011)	0.899 (0.016)
Q2	0.8	0.8 (0)	0.812 (0.062)	0.745 (0.077)	0.8 (0)	0.799 (0.001)	0.799 (0.002)	0.800 (0.004)	0.799 (0.008)	0.799 (0.013)
Q3	0.6	0.6 (0)	0.58 (0.140)	0.484 (0.15)	0.599 (0.001)	0.599 (0.003)	0.6 (0.006)	0.6 (0.011)	0.596 (0.023)	0.599 (0.031)
Q4	0.4	0.4 (0)	0.425 (0.069)	0.37 (0.066)	0.399 (0.001)	0.399 (0.002)	0.399 (0.006)	0.402 (0.01)	0.403 (0.028)	0.401 (0.035)
Q5	0.2	0.2 (0)	0.238 (0.181)	0.273 (0.183)	0.2 (0.002)	0.200 (0.006)	0.199 (0.01)	0.199 (0.021)	0.198 (0.044)	0.193 (0.062)
Q6	0.9	0.9 (0)	0.835 (0.104)	0.804 (0.129)	0.9 (0.001)	0.900 (0.003)	0.9 (0.005)	0.9 (0.009)	0.901 (0.021)	0.902 (0.032)

Table A-V: Extension of model to Knirps and Krüppel short-range repressors. Parameter estimations were done 1000 times for each formulation of the model (schemes 1-9) to find the R (repressor scaling factor) and C (cooperativity) parameters for Knirps and Krüppel, fixing quenching efficiency parameters to those of Giant. All schemes generated similar cooperativity values for the two proteins, and dissimilar repressor scaling factors. Parameter value (Standard deviation)

	Knirps				Kruppel			
Scheme #	ERROR	R	C1	C2	ERROR	R	C1	C2
1	0.007 (0)	1.3 (0.15)	0.88 (0.22)		0.01 (0.002)	37 (11)	1.1 (0.85)	
2	0.007 (0)	1.4 (0.19)	0.67 (0.28)	0.69 (0.19)	0.01 (0)	30 (6.4)	2 (0.82)	0.32 (0.31)
3	0.007 (0)	1.4 (0.18)	0.78 (0.22)		0.011 (0.002)	37 (11)	1.1 (0.97)	
4	0.007 (0)	1.4 (0.19)	0.7 (0.3)	0.5 (0.14)	0.011 (0)	29 (6.4)	2.3 (1.3)	0.31 (0.37)
5	0.007 (0)	1.3 (0.14)	1.1 (0.28)		0.009 (0.005)	33 (11)	1.6 (1.3)	
6	0.007 (0)	1.3 (0.16)	0.88 (0.23)		0.011 (0.003)	35 (12)	1.3 (1)	
7	0.007 (0)	1.5 (0.28)	0.54 (0.25)		0.019 (0.01)	29 (19)	1.7 (1.8)	
8	0.007 (0)	1.5 (0.2)	0.56 (0.28)	1.3 (0.23)	0.009 (0.001)	37 (8.8)	1.8 (1)	0.34 (0.3)
9	0.007 (0)	1.5 (0.2)	0.63 (0.3)	1 (0.17)	0.009 (0.001)	36 (8.7)	2 (1.7)	0.31 (0.35)