A Biological Approach for Obtaining Quantitative Data to Mathematically Model Transcriptional Regulation in *Drosophila melanogaster*

By: Kelsey Hansen

Department of Biology

South Hadley, MA 01075

May 2015

This paper was prepared with the guidance of Professor Craig Woodard of the Mount Holyoke College Biology Department

ACKNOWLEDGEMENTS

Most importantly I would like to thank my thesis and academic advisor, Professor Craig Woodard. His guidance over the past three years has been unrivaled. Without him this project would not be a possibility, as he took on the daunting task of becoming my thesis advisor and first reader when the choice was hardly his. Despite the lab that he already runs on this campus, and all of the students that he advises, and meetings he has on a daily basis, he was still kind enough to take me on and aid in the completion of this project.

This project also would not have been a possibility without the support of Robert Drewell, Jackie Dresch, who met me once over Skype and agreed to take me into their lab as an eager sophomore. Professor Dresch has been more than accommodating in helping me learn the ins and outs of MATLAB. Professor Jeff Knight has also been very helpful as he agreed to join my committee as the second reader.

The other member's of the Dresch-Drewell lab have been there for support throughout this whole process, especially Mehnaz Ali who gave me her undivided attention when something really didn't make sense or when this task seemed impossible to achieve.

Lastly, to my friends and family, this would not have happened if you hadn't been by my side through every step of this process. From the gentle encouragements, to the endless pep talks, you were able to put up with me over the past two years and I couldn't be more thankful to all of you.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Understanding the genetic control of development is both important and exciting because it has the potential to lead us to revolutionary discoveries that could provide advancements in numerous fields such as molecular biology, immunology, and the study of transcriptionally related diseases such as HPV. *Drosophila melanogaster* is a widely used animal model for studying the biological processes of development. It has a rapid maturation process and produces many offspring, making them easy to observe, especially during development. Early stages in *Drosophila* development and other multicellular organisms include critical events that lead to the differentiation of cells and eventually proper tissue and organ development in the adult stage. A complex network of genes, transcription factors and cis-regulatory modules controls these events. These modules could provide very important information about the regulation of transcription.

To determine the mechanisms underlying regulation of gene transcription, other investigators take apart predicted enhancer and transcription factor binding sites (a top-down model). In contrast, I am attempting to build from the bottom-up using a mathematical approach. I am using 10 different DNA constructs that were designed to represent single activator binding sites, as well as activator and repressor binding site pairs. My approach involves using MATLAB, along with basic biological assumptions and predictions, to model these complex binding events. The ultimate goal of this project is to create a model that can be used to compare biologically obtained data. Using this data we can then refine the mathematical model to more accurately represent transcription factor binding in the early stages of *D. melanogaster*.

INTRODUCTION

In molecular biology, transcriptional regulation is the process by which a cell can regulate the conversion of deoxyribonucleic acid (DNA) to ribonucleic acid (RNA) (transcription), playing a large role in gene activity (Griffiths *et al.,* 2012). In any multicellular organism, development from a single-celled zygote to a more complex adult body requires a very specific cascade of events. Events that contribute to tissue differentiation and organization must be controlled in a spatio-temporal manner.

Through a pattern of regulatory transcription factors deposited by the mother into the oocyte, the correct cell differentiation can be achieved in the developing embryo (Gilbert, 2000). Throughout the process of becoming a fully developed adult body, *cis*-regulatory modules (CRMs) are responsible for the expression patterns seen in the anteroposterior axis of the embryo. These activated modules will aid in driving expression of specific transcription factors. The cascade of transcription factors will eventually terminate with the homeotic, or *Hox* genes, which specialize in cell-differentiation (Sanson 2001).

It is these *Hox* genes that, with proper functionality, will ensure that the transformation from a larval body to an adult body happens properly. This is why gene expression must be very tightly regulated in the embryonic stage of development. Regulation of these genes eventually leads to proper adult body segmentation (Sanson 2001).

Drosophila melanogaster Life Cycle

Drosophila melanogaster is holometabolous, which means that before adulthood is goes through larval and pupal stages. It starts off as an egg that hatches in about 22-24 hours and it enters the first instar larval stage where it continuously eats. It maintains this stage for roughly 25 hours and then molts into a second instar larva where it continues to eat until it reaches its third instar larva stage in another 24 hours. In this stage it begins to climb its way out of its food so that it can prepare for its pupal stage. In about 30 hours the third instar larva becomes a pupa. Pupae are stationary and in the early stages can appear yellow/white. As it develops the pupa becomes darker and eventually metamorphoses into an adult fly. This happens within three to four days and the adult fly will emerge from its pupal case, known as an eclosion. Female eggs are not ready until about two days after this eclosion, but males are sexually active hours after. *Drosophila* animal models are great for molecular biological purposes for many reasons, but one of the most important is because of how rapidly they mature into adults and create new progeny (Griffiths, *et al*., 2012). A diagram of the *Drosophila* life cycle can be found in **Figure 1**.

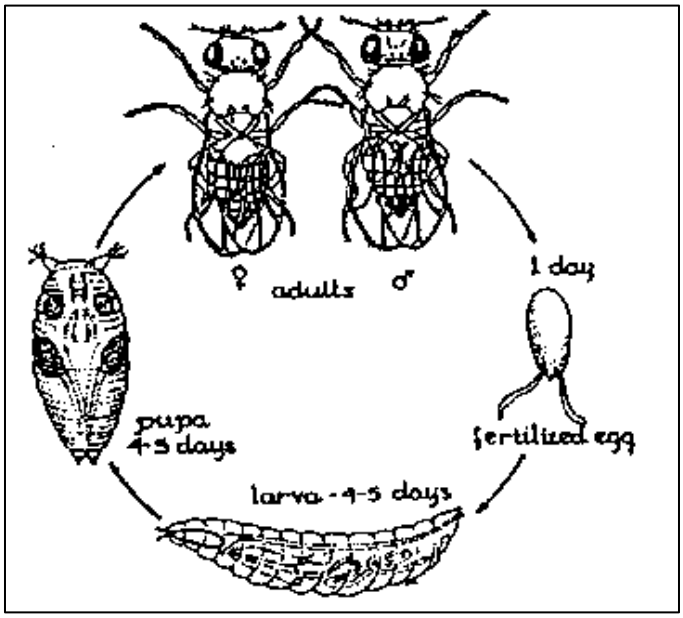


Figure 1. The life cycle stages of *Drosophila melanogaster*. The life cycle consists of and embryonic stage, three instar larval stages, and eventually pupariation and development into an adult fly body. (Figure taken from Mount Holyoke College Bio 210 Lab Manual 2007)

Drosophila melanogaster as a Model Organism

Drosophila melanogaster has played a key role in molecular biology and genetic studies for many reasons. First and foremost, it has significant applications in humans and other organisms not so suited for research purposes (Roberts 2006). Early researchers discovered that the genome of the fruit fly is highly susceptible to manipulation. For example in *Drosophila*, laboratory scientists can develop molecularly designed deletions, create targeted mutations and modify large fragments of DNA to carry out a large range of experiments (Venken and Bellen 2005).

Another key reason that *Drosophila* is so commonly used as a model organism is because of the duration of its life cycle. Within about 10 to 12 days the organism undergoes rapid transformation from embryo to adult. Once

it has reached the adult stage, it is able to produce a large number of offspring. Through classical genetics we can determine what proportion of these offspring will contain the mutation of interest.

Transcription Factors and *cis*-regulatory Elements

To ensure that proper adult body formation occurs, a cascade of regulatory genes is established by the deposition of regulatory transcription factors (*trans*-acting factors) by the mother into the oocyte (Gilbert 2000). These maternal transcription factors assist in the initiation of proper cell differentiation in the developing embryo. Throughout development *cis*-regulatory modules that are responsible for patterning of the anteroposterior (AP) axis become activated through interactions with the *trans*-acting factors (Dresch and Drewell 2012). This drives the expression of new transcription factors (TFs) in a cascade that results in the expression of the *Hox* genes, which control most of the cellular differentiation. The cascade of transcriptional events begins with mRNA transcripts encoding *bicoid and nanos*–deposited by the mother.

After being translated into proteins, BICOID and NANOS serve as activators to the gap genes that follow in the cascade. The gap genes include *hunchback*, *caudal* and *krüppel* primarily. There are also the secondary gap genes, which are *giant and knirps* (Sanson, 2001). [These genes all encode for proteins with the same name (HUNCHBACK, CAUDAL, KRÜPPEL,

GIANT and KNIRPS) which serve as transcription factors, playing a key role in the development of the *Drosophila* embryo.]

These genes are expressed in wide stripes across the developing embryo along the (AP) axis regulating the expression of the pair-rule genes such as *even-skipped*. These pair-rule genes are expressed in the embryo in a seven-stripe pattern, as shown by **Figure 2**, and many of the enhancers that control this expression are known, reinforcing the fact that *Drosophila* makes a good model organism for this project.

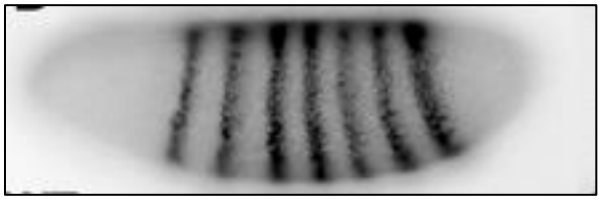


Figure 2. *Even-skipped* stripe development in a wild type embryo. *Even-skipped* (*eve*) is expressed in a seven-stripe pattern in wild type *Drosophila* melanogaster embryos. These stripes are established by the pair-rule genes and are then further refined by auto- and co-regulation among the pair-rule genes. Many of the enhancers that control the expression of these individual stripes are known. Further interactions between these pair-rule genes aid in establishing the segment polarity gene expression patterns which act with the HOX genes to determine the fate of each embryonic segment (Borok *et al*. 2010).

Figure adapted from Kim *et al*. 2011

There are several major mechanisms that are used to control and regulate the aforementioned developmental cascade. One of the more important mechanisms is the *cis*-regulatory modules, or CRMs. CRMs are regions of non-protein coding DNA that are in synteny to the genes they regulate. Being syntenic (or in more modern terms, in-cis) means that they are co-localized to the same genetic loci on the same chromosome within the

individual (Passarge *et al.,* 1999). There are several types of *cis*-regulatory sequences. Among them are silencers, insulators and enhancers. Enhancers are known for their ability to bind TFs in a sequence dependent manner (Borok *et al*. 2010) as well as their ability to make contact with the promoter of a gene (Ptashne 1986). Transcription factors can act as either activators or repressors. Activators can up-regulate the expression of a gene whereas repressors can down-regulate expression. Functional enhancers will generally be comprised of a mixture of activator and repressor binding sites, which allows for the distinct patterns noted in the developmental genes. By extension, these distinct patterns will allow for precisely regulated tissue differentiation into an adult fly (Borok *et al*. 2010).

Transcription Factor Functionality

Transcription factors work by binding to enhancers. This event can either up-regulate the transcription of the gene (activate) or down-regulate the transcription (repress). Activators and repressors work by binding to enhancers on the DNA. Each activator and repressor has a certain affinity for this enhancer sequence determining the frequency of binding. When an activator binds to an enhancer it will activate transcription by altering the chromatin structure through modulation of histone acetylation. On the other hand, repressors deacetylate histones and stabilize nucleosomes (Wolffe *et al.* 1997). Nucleosomes will create a closed environment and will inhibit transcription as depicted in **Figure 4**.The seven stripes seen during

embryogenesis can provide a model of the functionality of activators and repressors. In stripe two for example, the activators (BCD and HB) allow for the activation of *even-skipped* transcription, while the repressors (GT and KR) cause the sharp borders observed in this stripe as their function is to repress the expression of even-skipped (Stanojevic *et al.*, 1991). This is clearly depicted in **Figure 5**.

Transcription factors are a subcategory of DNA-binding proteins. These proteins contain a DNA-binding motif, which is the part of the protein that makes contact with the double helical DNA. In eukaryotes, the most common motif is the Zinc finger motif (Brown 2007). Hunchback and Krüppel, for example, are zinc finger proteins (Brody 2001). These motifs are generally about 25-30 residues long and contain two histidine residues and two cysteine residues. These residues coordinate a zinc atom and consist of a beta strand-turn-beta strand-turn-alpha helix complex when stably folded (Laity *et al*. 2001). (ADD FIGURE?)

Transcription Factors and Their Role in Transcription Initiation

Transcription initiation requires a TF to bind (by way of its DNA-binding motif) to the DNA on an enhancer by recognizing short sequences within its motif, also known as a transcription factor-binding site. A functional enhancer will have the ability to contact the promoter for a specific gene and recruit transcriptional machinery allowing for transcription to occur. Another method a functional enhancer could use is secondary activation using the

recruitment of coactivators. These may help to recruit chromatin modification enzymes (histone acetyl transferases – HATs) that promote successful transcription. See **Figure 4** for a diagram of enhancer activity.

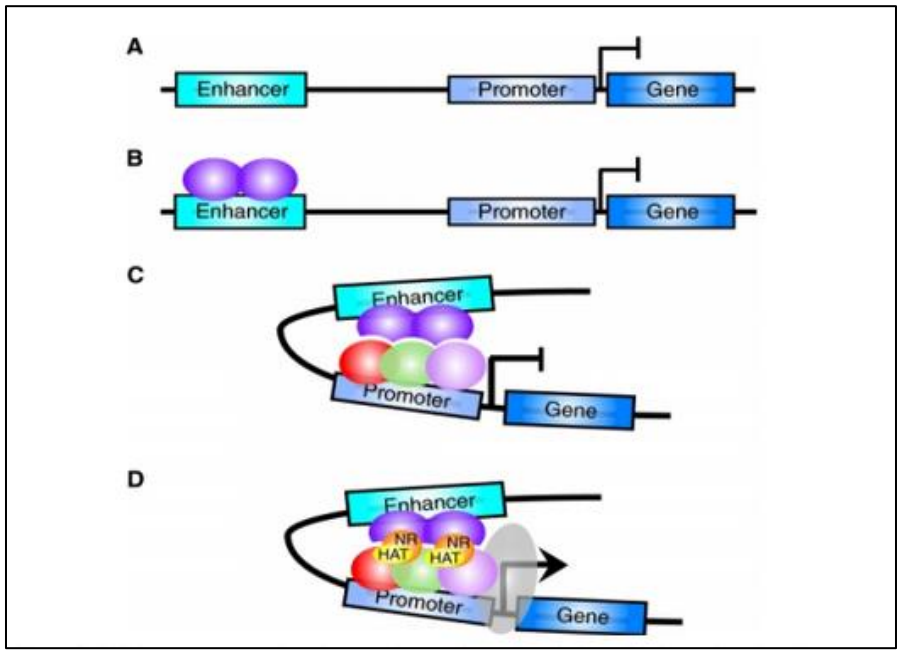


Figure 3. Diagram of transcription initiation. **A.** Enhancers are syntenic to the genes they regulate and are located in (usually) the introns of the genomic DNA. They are very rarely located in the exons due to their importance to gene transcription. **B.** Transcription factors, represented by purple circles, bind to the enhancer. **C.** Showing a major qualification of an active enhancer by having the ability to make contact with the promoter of the gene of interest. The enhancer is also in synteny with the promoter making this possible. Other types of transcriptional machinery are also being recruited by the enhancer-shown in red, green and pink. This machinery makes it possible for transcription to occur. **D.** The recruitment of HAT can also be crucial to the regulation of transcription as it acts as a chromatin modification enzyme. The recruitment of a nuclear receptor coactivator can also be important as it directly acts with transcriptional machinery as well.

Figure from Borok *et al.* 2010

There are several different mechanisms proposed for the up-regulation of genes. Activators can act as recruiters of RNA polymerase II (**Figure 3C** Kim and Lis 2005). They can also induce an open chromatin configuration around the promoter of the target gene (Adkins *et al*. 2006). This can be

achieved by HAT activity, which adds acetyl groups to the histone tails. There are times when an open chromatin conformation already exists around the promoter. In this case, nuclear receptor coactivators can be recruited to recruit RNA polymerase II

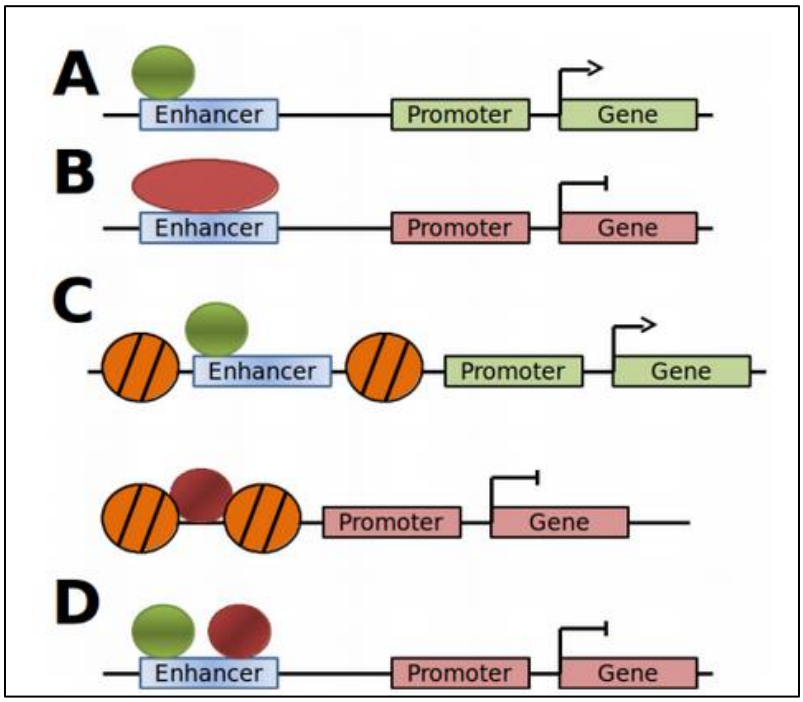


Figure 4. Diagram of repressor mechanisms. **A.** This depicts a lack of repressors bound to the enhancer. Only an activator is bound to the enhancer allowing for transcription of the target gene to occur. **B.** The red oval represents a repressor that bound to the enhancer in place of the activator. This repressor is not allowing for the transcription of the target gene. **C.** Repressors that are depicted in red circles act in the manner of chromatin remodeling. This chromatin remodeling requires the recruitment of histone deacetylation complexes (HDACs) that will cause a closed chromatin environment not allowing for transcription to occur, whereas nucleosomes (orange circles) cluster more closely together and prevent activator binding. **D.** This depicts a case of long-range repression. This long-range repression is attempting to show that even though an activator is bound, the long-range repressor is preventing transcription of the target gene.

Figure adapted from (Brown 2013)

Repressors act very differently than activators do. Repressors can act to disrupt activator binding or they can directly influence the promoter region of the target gene. If the repressor is acting to disrupt activator binding, this is known as short-range repression and can cause competitive binding (activator vs. repressor) or quenching. Quenching can also be described as incomplete repression. In this case there would still be slight expression of the gene of interest, but the expression would be much less robust (Li and Arnosti 2011).

Eve Stripe 2 Formation in the Developing *Drosophila* Embryo

As previously mentioned, if we look specifically at the *even-skipped* (*eve*) gene, whose mRNA transcripts are observed as a tightly regulated, characteristic, seven-stripe pattern across the developing embryo, there are defined gradients of certain transcription factors. These gradients show examples of activators and repressors interacting within an organism. These transcription factor gradients are key to the proper organization and segmentation in the adult *Drosophila* body. Eve stripe 2 is a highly studied area of *Drosophila* development. Studying *eve*, scientists were able to determine that multiple binding sites and *trans*-acting regulatory factors are required for proper enhancer function (Ludwig and Kreitman 1998). Looking closer at eve stripe 2 enhancer, we can determine the DNA sequence and bioinformatically predict binding sites for transcription factors. These can then be compared to the analyses of the concentration gradients for transcription

factors that arise in the early *Drosophila* embryo, providing evidence for why the boundaries of the eve stripe 2 (*S2E)* are so sharp.

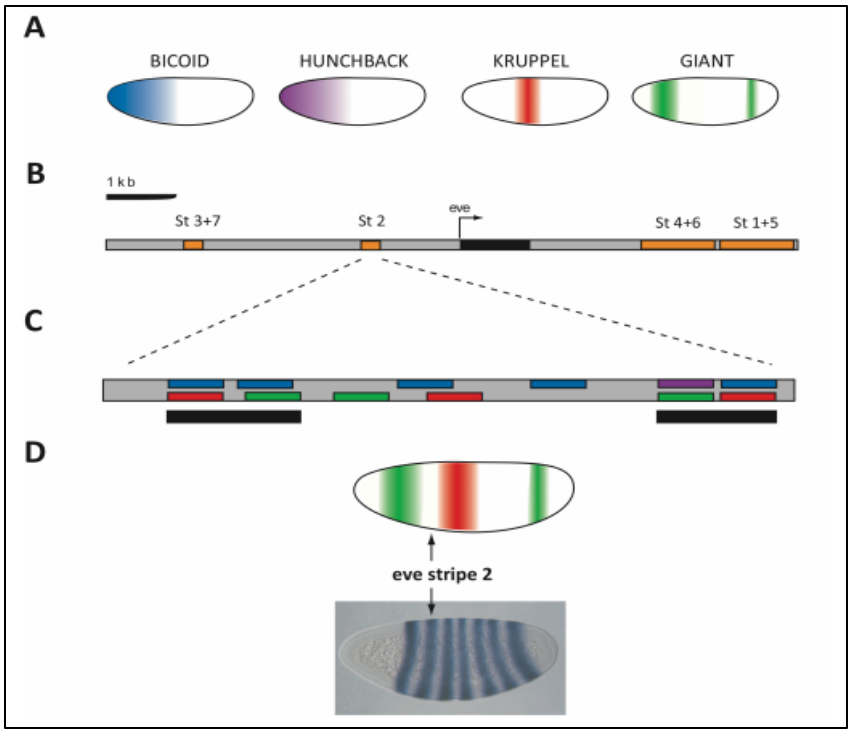


Figure 5. Activation of *eve* expression and the transcription factors involved. **A.** Transcription factor gradients and localization patterns for BCD, HB, KR and GT respectively. These are the four major transcription factors involved in the formation of S2E. BICOID and HUNCHBACK are activators of the *even-skipped* stripe 2 enhancer. KRUPPEL and GIANT are the repressors. **B.** Depicts a map of the *eve* genomic locus. The orange bands represent the enhancers that are responsible for the (black) expression that is shown. **C.** This breaks down the *eve* locus even further to just the stripe 2 enhancer. The colored bands are coordinated to each individual transcription factor binding site located within the stripe 2 enhancer. The two black bars are indicating a cluster of binding sites that are thought to be particularly important to the activity of this enhancer. **D.** This image is depicting the defined second stipe in a real *Drosophila* embryo. These defined lines are established by high concentrations of GIANT at the anterior boundary and KRUPPEL at the posterior boundary.

Image adapted from Dresch and Drewell 2012

Quantitative Transcription Factor Data

The data obtained by the Luengo lab in the *Drosophila* Transcription Network Project based out of the University of California Berkeley, provides quantitative data for many different transcription factor concentrations in a developing embryo. These values are given at multiple time points for each transcription factor across 52 nuclei that span the anterior to posterior of the embryo. When all of these transcription factor gradients are plotted on the same graph across a single time point, as shown by **Figure 6**, the data will show (in a broader context) how these transcription factors work together at this time in development. There were many different time points taken in this experiment, each one taken ten minutes apart. These tight time points in development later make our thermodynamic assumptions acceptable as we don't expect much change to occur in that time frame.

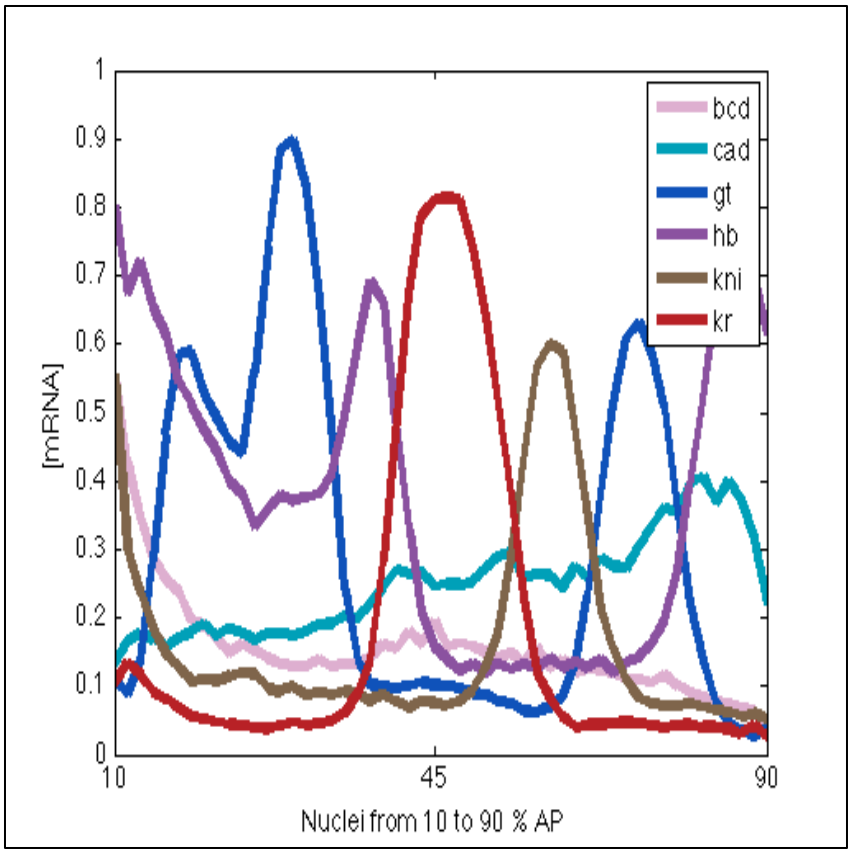


Figure 6. Transcription factor concentration gradients. Graph of quantitative mRNA concentration values of 6 transcription factors across 52 nuclei in a developing embryo from anterior to posterior during the second time point of measurement. This graph shows the localization of these transcription factors during this stage of embryogenesis (Luengo *et al*. 2006). The Berkeley Drosophila Network Project measured this data for the purposes of developing mathematical models.

Mathematical Modeling and its Biological Relevance

Mathematical modeling and bioinformatics analysis has become very prominent in biological research. For example, bioinformatics analysis has revealed that the transcription factor binding sites within S2E are not very well conserved between orthologs (Ludwig *et al*. 2000). It has also shown that of the seventeen known binding sites only three are completely conserved at sequence level between the different *Drosophila* species. This bioinformatically derived genetic information guides the future research for

why these sites were so important for transcriptional regulation throughout evolution and time.

Bioinformatics has become a crucial part of modeling genetic transcription and its regulation. A very important tool called PATSER, invented by Jerry Hertz (Helden, 2003) was utilized to bioinformatically predict putative transcription factor binding sites. Another tool EvoPrinterHD was used to determine conserved regions within the enhancers (Yavatkar et al., 2008). Adam Brown, a Harvey Mudd student who started this project, ran the bioinformatic tests to find these sites. He also tested each putative functional motif that was strategically selected through thresholding and creating statistically significant cut-off values. Compiling all of the information regarding bioinformatically predicted binding sites as well as previous knowledge about short-range repression (100 base pairs is the longest distance that can produce this result), putative functional motifs (gBlocks) were designed to fit these required parameters using the design laid out in **Figure 7**.

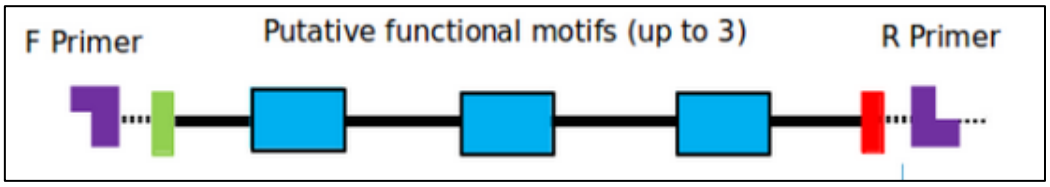


Figure 7. Generic gBlock insert design. The two purple L shaped blocks represent the forward and the reverse primers needed to amplify the DNA using PCR. The red and green blocks represent very distinct restriction enzyme sites (*Not* I and *Xho* I). These allow for traditional cloning processes. The black bars are the spacer regions (comprised of 100 base pairs).These spacer regions are placed between each of the blocks to allow for combinatorial testing of these motifs without repressor interference between them by means of short-range repression. The dotted black lines are subsections that ensure exonuclease activity does not disrupt any of these functional motifs (Brown, 2013).

After determining the motifs to be tested 10 different constructs were designed (**Table 1)**. These constructs contain 5 different activator and repressor pair motifs, 3 activator only motifs and two different controls. One of the controls has a 2F2K region, which is known to give expression in early embryos. There is also a negative control, which should show no expression in the embryos validating a correct in-situ procedure and the results obtained from the other constructs.

Table 1. Construct and descriptions

Construct Name	Description
BCDKR	Motif Flanked by spacers
HBGT	Motif Flanked by spacers
HBKR	Motif Flanked by spacers
CADGT	Motif Flanked by spacers
BCDKNI	Motif Flanked by spacers
BCD	Consensus BCD sequence flanked by spacers
HB	Consensus HB sequence flanked by spacers
CAD	Consensus CAD sequence flanked by spacers
SPACER_2F2K	2F2K region flanked by spacers
SPACER_NEGATIVE	2 concatenated spacer regions

Table from (Brown, 2013)

These strategically designed constructs were developed to serve as a bottom-up strategy for studying transcriptional regulation. The intent of these constructs was to later insert them, using molecular biology techniques, first into a plasmid vector containing a *lacZ* reporter gene, and then create a line of *Drosophila* that contain these constructs. After the insertion, embryo collections can be done and in-situ hybridization performed. The in-situ will stain the embryo wherever the *lacZ* has been transcribed, which will hopefully

be where these transcripts are located. Images of these embryos can then be run through a pipeline for determining the quantitative values of expression, which can then be compared to the mathematically predicted outputs for the same constructs.

Modeling the transcription of genes in *Drosophila* could provide great insight on transcription in other organisms. Modeling in this case means that there is an equation or a system of equations that can describe the physical aspects of a biological process (Dresch and Drewell 2012). Developing these equations allows mathematicians and biologists to work closely and efficiently by making parameter predictions with the ability to run numerous tests at a time on a computer and then test those results for accuracy in a biological or experimental manner. This can even be accomplished in the reverse order. Biologists can use what they already know and with the help of a mathematician, create a mathematical representation of these data.

Thermodynamic Modeling

Thermodynamic modeling is a commonly used type of math modeling. It can also be known as the fractional occupancy model (Dresch and Drewell 2012). The idea behind this type of model is that it will create an output of predictions based on "successful" states as a fraction of all possible states for that particular construct. This is a good model design for transcriptional regulation. This model can be used to predict mRNA concentrations and therefore gene expression if we make the assumption that the mRNA

concentration is proportional to the outcome of predicted expression and vice versa. This then allows us to compare both of sets of data, predicted and biological, and determine the true values for these constants.

To accomplish this we must first determine how this fractional occupancy model should be represented. Assuming that there is thermodynamic equilibrium, we must take into account the DNA binding motif, or the enhancer, M, and the different transcription factors, TF, that have the ability to bind to this segment of DNA. The transcription factor can be either bound, or unbound to this enhancer, as seen here:

$$M + TF \underset{K_U}{\overset{K_B}{\rightleftharpoons}} M \cdot TF$$

Where K_B represents the constant rate of binding of TFs binding to the motif, and K_U represents the constant rate of unbinding or dissociation for the TF dissociating from the motif.

In the thermodynamic equilibrium model, we must assume that the binding and unbinding rates of these transcription factors are equal, and that both the products and reactants should be constant and the rate of change is equal to zero. We can then compare the ratio between the binding affinity constant and the dissociation constant, leading to the formation of a probability equation.

$$P = \frac{[M \times TF]}{[M] + [M \times TF]} = \frac{\frac{[M \times TF][TF]}{[M][TF]}}{\frac{[M][TF]}{[M][TF]} + \frac{[M \times TF][TF]}{[M][TF]}} = \frac{K_{TF}[TF]}{1 + K_{TF}[TF]}$$

This equation has the ability to output the probability that any transcription factor will be bound to any motif (enhancer). P represents the probability output, M, the motif, which in this case represents an enhancer. M·TF represents a transcription factor bound to a motif and TF represents one more of these particular transcription factors. This equation shows that the probability of transcription occurring is equal to the probability of being bound divided by all possible states (bound and unbound). The value of 1 represents the empty state where nothing is bound to the enhancer. The K parameter represents the binding affinity constant of each respective transcription factor, which has not yet been determined.

The transcription factor in this equation can act as either an activator or repressor. Thinking in terms of probability, the probability of expression happening is when a TF activator is bound (successful states) over all of the possible states for that TF. This gets much more complicated when both activators and repressors are present. The successful states can be defined as a bound activator. An unsuccessful state can be defined as a bound repressor, and if they are both bound to the motif, it will be considered successful. This equation represents the presence of only activator transcription factors, with no repression:

$$[mRNA] \mu \frac{K_A[A]}{1+K_A[A]}$$

A represents an activator, and there are no other parameters involved in this model.

When considering both an activator and a repressor, two more parameter values must come into play. You must now take into consideration the binding affinity value for the repressor (R) and the activator (A). Not only can a repressor fully repress transcription, but it also has the ability to slightly quench transcription. In this case we should still see expression, but it appears much less "robust" when compared to the single activator state. We can assume that Q has a multiplicative effect on the ability to repress transcription, and on the ability of the enhancer (Fakhouri *et al.* 2010;Sherman and Cohen 2012). The equation that represents the probability of expression level under this condition is as follows:

$$[mRNA] \mu \frac{K_A[A] + QK_AK_R[A][R]}{1 + K_A[A] + K_R[R] + K_AK_R[A][R]}$$

This represents an enhancer with two binding sites, one for an activator and one for a repressor. K_A is the binding constant for the activator; K_R is the binding constant for the repressor. The Q term represents the quenching coefficient where 1-Q represents the quenching ability of the repressor. When Q=0 we expect to see no expression as it will fully repress, and when Q=1 we expect to see full expression, as this is a poor repressor.

Quenching is not the only factor that we must take into account. Another possibility is that there is an enhancer, again with two binding sites, but in this case the binding sites are for two activators. If two activators are bound there is a potential for cooperativity. In this case, there are many successful states if we again, assume that a successful state is only one activator bound, meaning that the only unsuccessful state is when there is

absolutely nothing bound to the enhancer. The expression equation would look like this:

$$[mRNA]\mu \frac{K_{A1}[A1]+K_{A2}[A2]+CK_{A1}K_{A2}[A1][A2]}{1+K_{A1}[A1]+K_{A2}[A2]+CK_{A1}K_{A2}[A1][A2]}$$

This is an expression output equation with two activators and a cooperativity factor. A1 and A2 represent the two activators that have the potential to be bound to the DNA binding motif. Each of these activators have their own binding constant represented by their respective Ks. The variable C represents the cooperativity constant between A1 and A2, which is also multiplicative. This is a model that I did not run, but should be run in the future when more is known about the binding affinity constants for certain activators (Dresch and Drewell 2012).

Hypothesis and Goals

As we have seen, this equation can be built up and manipulated to create a very complex mathematical idea of transcription, but transcription on its own is a complicated biological process. The original goal of this research was to obtain quantitative values for the unknown parameters using molecular biology techniques, but due to unforeseen circumstances, the goal is to now run these models with expression data obtained from the University of California, Berkeley. This expression data makes modeling the above equations possible. Therefore, obtaining values for those unknown binding constants, quenching factors and cooperativity coefficients is also possible. Once the numerical predictions of these parameters has been obtained, they

can be tested in a molecular biology setting, where through experimentation, it will be feasible to confirm the quantitative data and either accept or reject the thermodynamic models, which will serve as my hypothesis.

Obtaining quantitative values will eventually lead to refining these models and being able to test these results across an even broader scale of transcription factors and motifs. In summary, the goal of this project is to use both molecular biology techniques and mathematics to obtain comparable data to both refine the already developed mathematical model as well as gain more knowledge and insight about how transcriptional regulation works in a more numerical sense.

MATERIALS AND METHODS

Molecular Biology

We have attempted to use DNA constructs to create new lines of flies to analyze the different expression patterns we are hoping to see with these constructs. This process began with obtaining plasmid vector DNA. This DNA would eventually be used to insert the gBlock constructs to be microinjected into flies. To obtain enough DNA of the cloning vector placZ.attB (originally provided by Harvey Mudd student Matthew Borok in 2007) the vector was plated on LB plates containing 150 mM ampicillin and incubated at 37°C for 14 hours. The colonies that grew were picked from the plates and grown in 1.5 mL of LB broth containing 50 mM of ampicillin overnight for 14 hours at 37°C with agitation. The cells were then centrifuged at 11,000 rpm for 1 minute and the LB broth was removed. The placZattB vector was isolated using the Zyppy miniprep kit.

To allow for ligation of the motif constructs into the vector containing the reporter gene, 100 µL of vector were digested overnight with 4 µL of each enzyme (*Not* I and *Xho* I). 12 µL of 10X 3.1 NEB buffer was added to the digest as well. This was incubated overnight at 37°C. After about 14 hours, 5 µL was electrophoresed on an agarose gel containing 1 µL of 10mg/ml ethidium bromide. After confirming on the gel that the enzymes cut, the vector DNA was dephosphorylated using 9.3 µL of phosphatase buffer and 1µL of Antarctic Phosphatase (NEB). This was incubated at 37°C for 15 minutes and then at 70°C for 5 minutes to kill the enzyme.

The rest of the DNA was electrophoresed to be prepared for purification. The purification step was carried out using the Zymoclean Gel DNA Recovery Kit. The success of the purification was confirmed by electrophoresing 2µL of purified DNA. The PCR products made in the previous semester were restriction digested. The restriction digestion reaction mixtures contained 15 µL of DNA (using only the positive control 2F2K product for now), 1 µL *Not* I, 1µL of *Xho* I, 2 µL of 10X NEB buffer and 1 µL of distilled water and left at 37°C for one hour. This was then purified using the DNA Clean and Concentrator -5 kit from Zymo Research.

Using the band intensities taken from the gel and the ladder comparison I determined how much insert and how much vector was needed to create a ligation that would yield 3:1 ratio as desired (3 insert colonies to 1 vector colony). This is determined by a mathematical method:

$$\frac{(ng\ of\ vector * kb\ size\ of\ insert)}{(kb\ size\ of\ vector)} * molar\ ratio\ of\ \frac{insert}{vector}$$

This equation can be used to determine the amount of insert vs. vector when attempting to create ligation colonies. (Typical insert/vector ratio is 3:1)

A control was used that contained vector only. Eventually when these were plated I hoped to see 3 ligation colonies on a plate to every 1 vector only colony.

Table 2. Ligation table

	Ligation	**Vector Only**
Vector DNA	4µL	4µL
Insert DNA	1 µL	-
T4 DNA Ligase	1µL	1 µL
10X Buffer	1µL	1 µL
Distilled Water	3µL	4 µL
TOTAL	10 µL	10 µL

This ligation was allowed to sit at room temperature for 2 hours and then overnight at 4°C.

The ligation was introduced into bacteria using Electomax electrocompetent DH10B cells stored at -80°C. 40 µL of cells were put into each of the two cuvettes. 2 µL of vector DNA were added to one of the cuvettes while 2 µL of the ligation were added to the other. These were then placed into an electroporator with voltage set to 2.00 volts, capacity set to 25 µF, resistance set at 200 Ohms to get a reading of time constants. The constant for vector only was 4.2 and the ligation was 4.0 (Potter and Heller 2010). Using a glass pipette, 1 mL of LB broth was added to the cuvettes to recover the cells. These were placed into a culture tube and were incubated for one hour at 37°C for full recovery. These were then plated on LB agar plates with ampicillin. Four plates were made: 200 µL vector only, 400 µL vector only, 200 µL ligation and 400 µL ligation.

Mathematics

Using MATLAB, a software technology allowing for designing and running mathematical models in a timely manner, I was able to write the code to run models for two situations. The two situations we would be looking at were a single activator, and an activator and repressor pair. The equations described in the introduction section were the basis of these models when thinking about writing the code. The codes used for a single activator and an activator and repressor pair are shown in **Figures 8** and **9** respectively.

```
for i=1:length(Conc)
    M(i)= k*Conc(i)/(1+(k*Conc(i)));
end;
```

Figure 8. Mathematical code representing the single activator equation. This code is written in a way that it will give an output value (M = expression) for any input data value (i). Here k represents the binding affinity of the activator, Conc represents the concentration of that activator (obtained from UC Berkeley data), and 1 represents the unsuccessful state (empty state) of binding by the activator. This code is run through MATLAB with all of the activator concentrations (HB, BCD and CAD) individually. This code will output the value for expression and can be used to find a rough data value for the binding affinity of the activator.

```
for i = 1:length(ConcA)
    M(i)=
    (KA*ConcA(i)+(KA*K
    R*Q*ConcA(i)*ConcR(i
    )))/(1+KA*ConcA(i)+K
    R*ConcR(i)+KA*KR*C
    oncA(i)*ConcR(i));
end;
```

Figure 9. Mathematical code representing the activator and repressor pair equation. Similarly to Code 1, this code is written in a way that it will give an output value (M= expression) for any input data value (i). In this code, KA represents the binding affinity of the activator, KR represents the binding affinity for the repressor, Q represents quenching ability of the repressor present, the ConcA represents the concentration of activator and ConcR represents the concentration of the repressor (obtained from UC Berkeley data). This code is run through MATLAB with all of the activator and repressor pairs (BCD KR, HB KR, HB GT, CAD GT and BCD KNI) individually. This code will output a value for expression and it can be used to determine the most ideal or best-fit value for Q as well as the binding affinities for both the activator and the repressor.

When running these codes through MATLAB, there are a significant number of unknown values; therefore it is necessary to make assumptions and predictions about the best values to assign to these unknowns to receive a biologically and mathematically sound value for M (expression). To start, I picked appropriate values for each of the different parameters in the single activator equation. I decided to run the equation in MATLAB with binding affinity values (K) between 10 and 100 increasing in increments of 10. MATLAB outputs a scatter plot for each output (M) value. This can later be used to compare to the in-situ hybridization results from the Berkeley data.

The shape and predicted expression values were compared to validate the mathematical model. The in-situ hybridization of the single activator constructs designed by Adam Brown would also be helpful in predicting which values from the mathematical predictions are the most accurate.

After noticing that the binding affinity values for my original prediction were much too high to represent actual biological data (Berkeley's), I decided to run MATLAB again and try running the model with much lower binding affinity values to see if I could more accurately represent the data from Berkeley. This time I ran the single activator code assigning binding affinity values from 1 to 9 by increments of 1. This can later be compared to the in-situ hybridization of the single activator constructs and a definitive value can be chosen for the binding affinities of these transcription factors.

After running the predictions for the single activator model, I moved on to the (five) activator and repressor pairs. These pairs are HB/KR, BCD/KR, HB/GT, CAD/GT and BCD/KNI. For each of these pairs I chose to assign values to the binding affinities of both the activator (KA) and the repressor (KR) that increased by increments of 10, just as before. I also chose to run each KA/KR pair with Q values between 0.1 and 1 increasing by increments of 0.1. Q represents the percent of time that the repressor is not repressing, while q (q=1-Q) represents the time that the repressor is repressing. In this context I will be using Q. Each activator and repressor pair had 100 graphs associated with it. Each of these graphs contains a certain KA/KR

pairing showing 10 different predicted expression values across the embryo for different values of Q.

RESULTS

Molecular Biology

After multiple attempts to get the enzymes (*NotI* and *XhoI)* to cut the placZ.attB vector DNA and running 1 µl on a gel to confirm the enzymes cut (**Figure 10**), I was able to run a full double digest (**Figure 11**) and cut the DNA band out of the gel for purification. This DNA was dephosphorylated for use later in the ligation. The dephosphorylating step prevents vector only ligations from ligating back together at a significant number. Once it was dephosphorylated the full amount of digest was run on a gel for purification. Once the full digest was purified it was sent to the ligation process. This process unfortunately never worked over the summer session of research. Thus far, we have not been able to get the desired result to build the constructs to get them injected into flies.

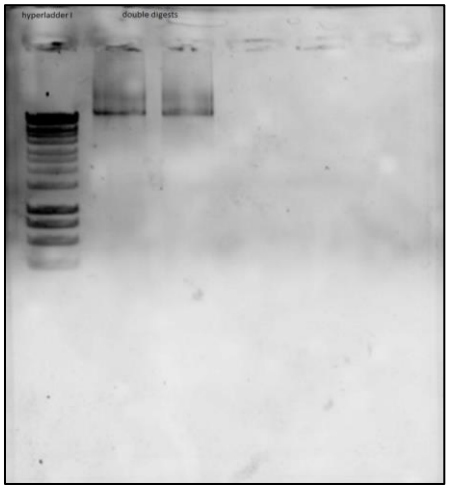


Figure 10. Gel confirmation of overnight digest of plasmid vector. The first lane contains the comparative ladder named HyperladderI. This ladder allows us to determine the size and the intensity of the band of the vector and confirm that it is the DNA that we were trying to obtain. The next two lanes are the full digest. Due to the distinct bands that we can see that have sharp edges, we can confirm that the digest has cut and proceed to the dephosphorylation process.

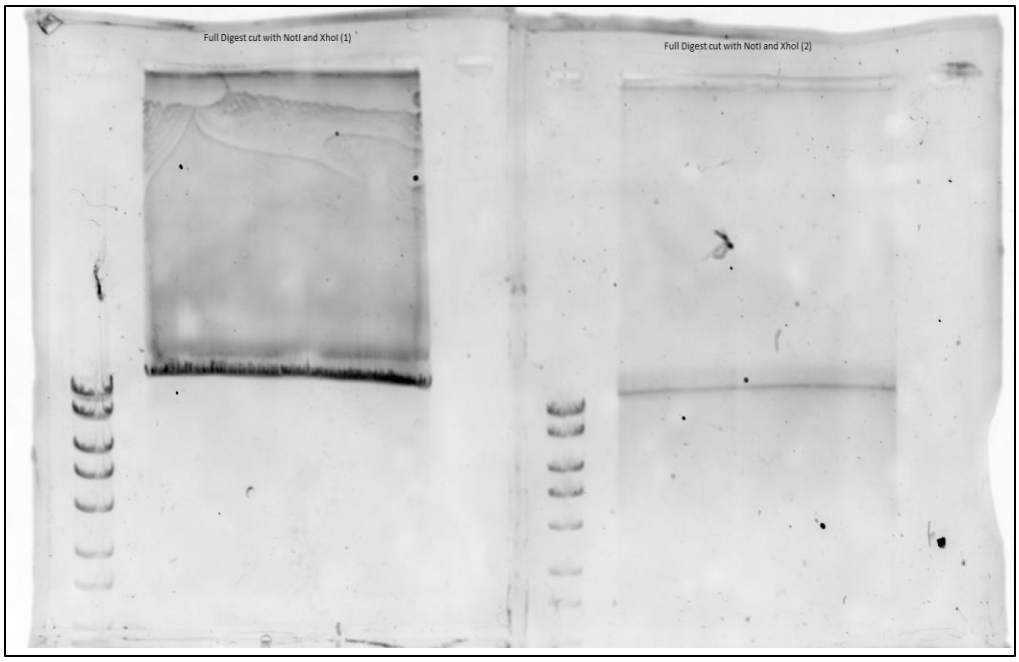


Figure 11. Full overnight digest. Full digest run on agarose gel to be cut out and purified. The first lane contains the comparative ladder HyperladderI allowing for the confirmation that the DNA is vector DNA using the known lengths of both bands. Both of these gel images contain vector DNA that were cut, confirmed and dephosphorylated. Running this on the gel allows for the beginning of the purification process. The bands of DNA are cut out using a UV light and a straight blade razor.

Unfortunately after all of the DNA was purified in both the constructs and the plasmid vector, when attempting to insert the constructs into the vector DNA in bacteria, the only colonies that we observed on the plates post-ligation were vector only at a very minimal amount. There are many reasons that these ligations could have failed to produce the necessary result. Molecular biology can be extremely difficult. Human error could play a huge role in the reason that some of the experiments did not work the first time, or at all. There are many steps that could be missed, messed up or for some reason just did not work due to errors beyond my control.

Mathematics

Although the biology portion of this project failed to yield many results, the opposite can be said for mathematical portion. Running my MATLAB code with varying parameters provided me with interesting insight to the biologically potential range of numerical values for both parameters. I ran three single activator constructs with my single activator code (BCD, HB and CAD). Each of the predicted expressions for these constructs are shown in **Figures 12, 13** and **14**).

Single Activator State

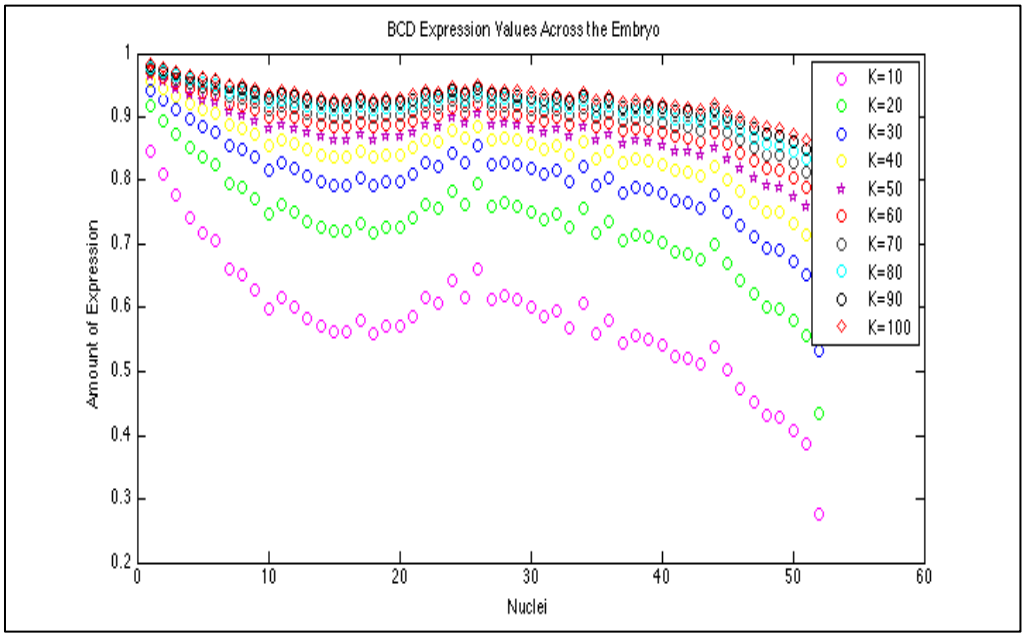


Figure 12. BCD single activator state. A range of expression values corresponding to an enhancer with a single BCD binding site across the 52 nuclei of the embryo from anterior to posterior during early development with varying binding affinity constants. The binding affinity constants range from 10 to 100 in increments of 10. The y-axis represents the percentage of expression in a range from 0-1. As K increases, so does the amount of expression.

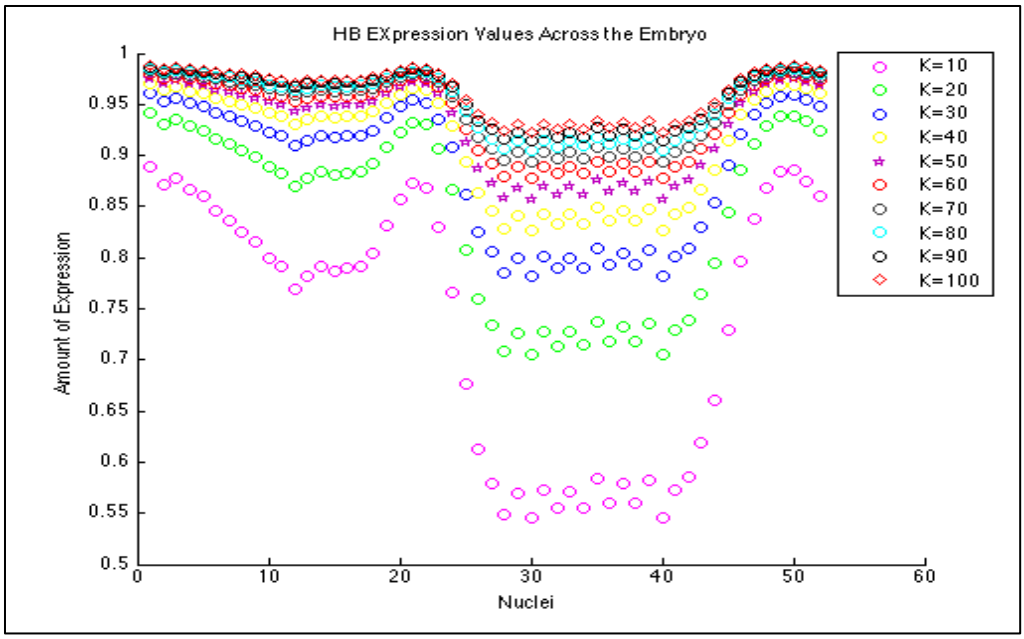


Figure 13. HB single activator state. A range of expression values corresponding to an enhancer with a single HB binding site across the 52 nuclei of the embryo from anterior to posterior during early development with varying binding affinity constants. The binding affinity constants range from 10 to 100 in increments of 10. The y-axis represents the percentage of expression in a range from 0-1. As K increases, so does the amount of expression.

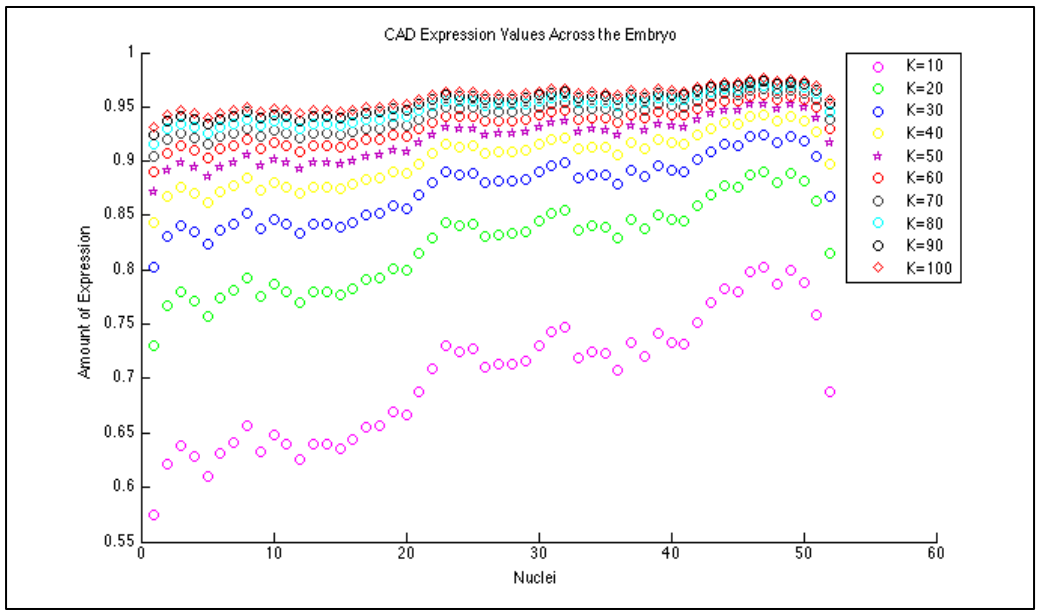


Figure 14. CAD single activator state. A range of expression corresponding to an enhancer with a single CAD binding site values across the 52 nuclei of the embryo from anterior to posterior during early development with varying binding affinity constants. The binding affinity constants range from 10 to 100 in increments of 10. The y-axis represents the percentage of expression in a range from 0-1. As K increases, so does the amount of expression.

These results match up well with **Figure 6** following the transcription factor mRNA concentration patterns obtained by the Berkeley *Drosophila* Transcription Network Project. This provides a good check that the code is running properly, and that the mathematics behind this idea is correct.

Activator and Repressor Pairs

The next step was to try the five different activator and repressor pairs. This model had three parameters in total: the binding affinity constant for the activator, the binding affinity for the repressor and the quenching value. The binding affinities were each increased by increments of 10, and Q values were tested from 0.1-1. (Q is the percentage of time that the repressor is NOT repressing. Therefore 1-Q=q and q is the amount of time the repressor is repressing. When Q is approaching 1 it is not a very good repressor and the expression values will increase.)

There were five total pairs to test, and because each binding affinity increased by 10 for each pair, there were a total of 100 graphs with 10 different Q values per graph for an overall total of 500 graphs. After sifting through these, I have decided to show representative graphs from each pair that comes directly from the middle of the possible range of data in **Figures 15, 16, 17, 18 and 19**. (See the Appendix for more of these graphs).

HB/KR Pair

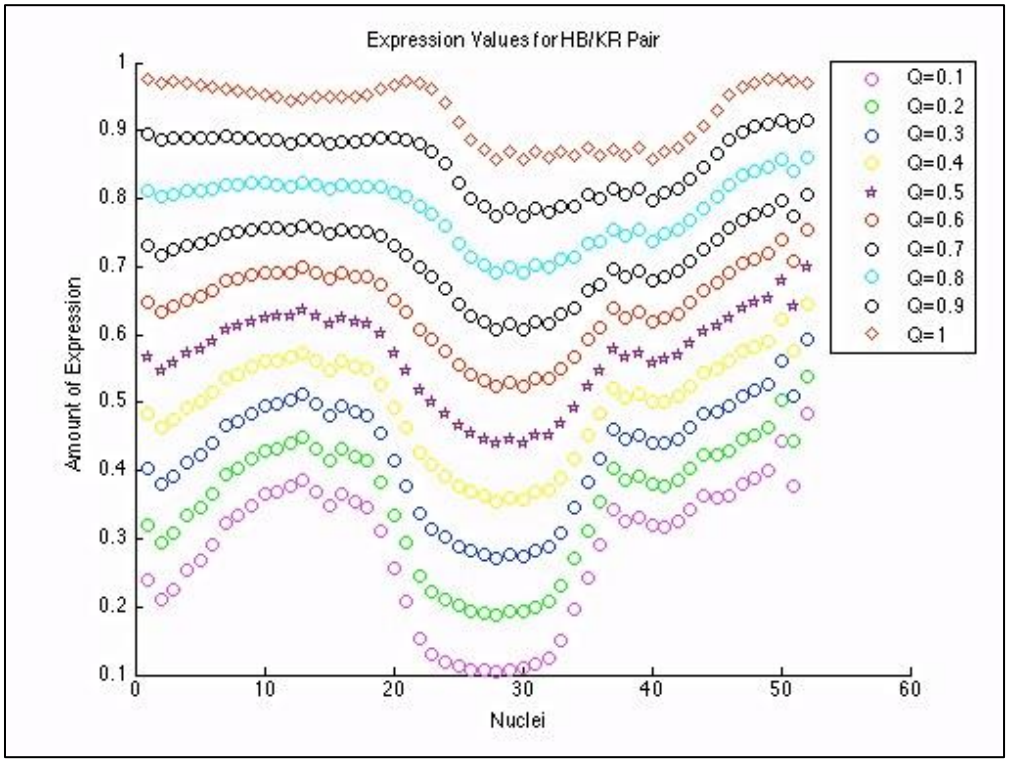


Figure 15. A predicted range of expression values for an enhancer containing both a HB and KR binding site across the 52 nuclei of the developing embryo from anterior to posterior. The binding affinity constant for the activator is: 50. The binding affinity constant for the repressor is: 50. The y-axis of this graph represents the fraction of expression between 0 and 1. The quenching values (Q) range from 0.1-1. These values may not be biologically accurate, but they serve as a good representation for the large amounts of data. This graph shows a clear increase of expression at the anterior of the embryo between 0 and 20 nuclei and again between 35 and 52 nuclei. This shows that our repressor is most active in the middle of the embryo. This matches up with the concentration of KR being heaviest in the middle of the embryo.

BCD/KR Pair

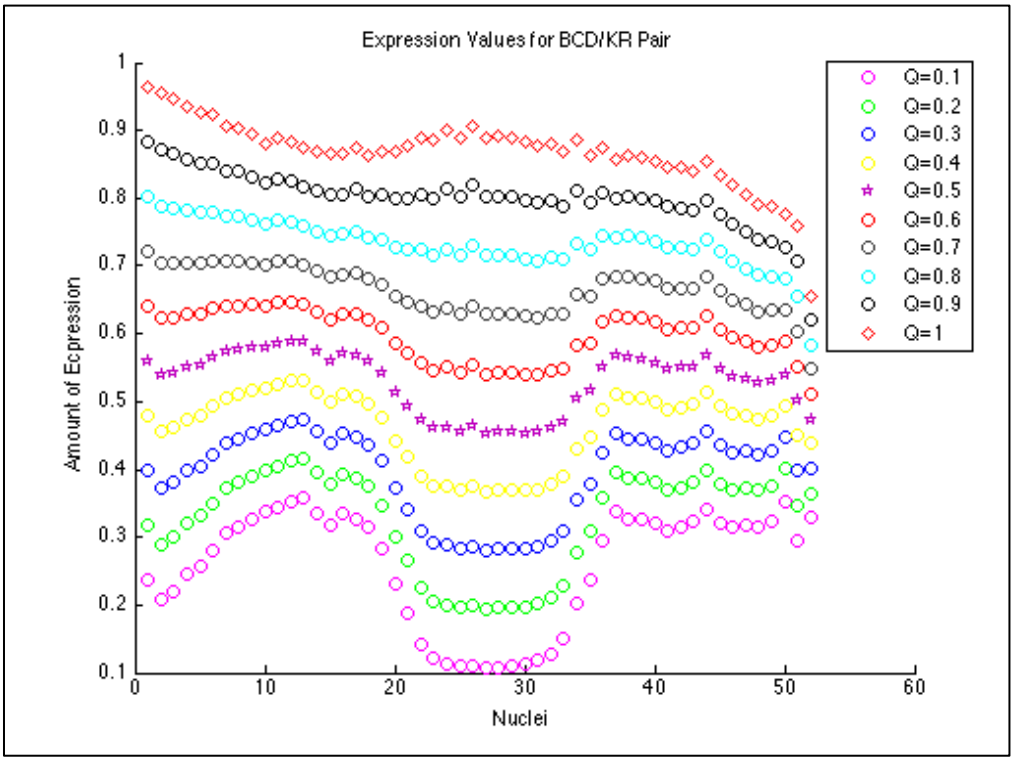


Figure 16. A predicted range of expression values for an enhancer containing both a BCD and KR binding site across the 52 nuclei of the developing embryo from anterior to posterior. The binding affinity constant for the activator is: 50. The binding affinity for the repressor is: 50. The quenching values (Q) range from 0.1-1. The y-axis of this graph represents the fractional probability of expression between 0 and 1. This graph shows a clear increase of expression at the anterior of the embryo between 0 and 20 nuclei and again between 35 and 52 nuclei. This shows that our repressor is most active in the middle of the embryo. This matches up with the concentration of KR being largest in the middle of the embryo.

HB/GT Pair

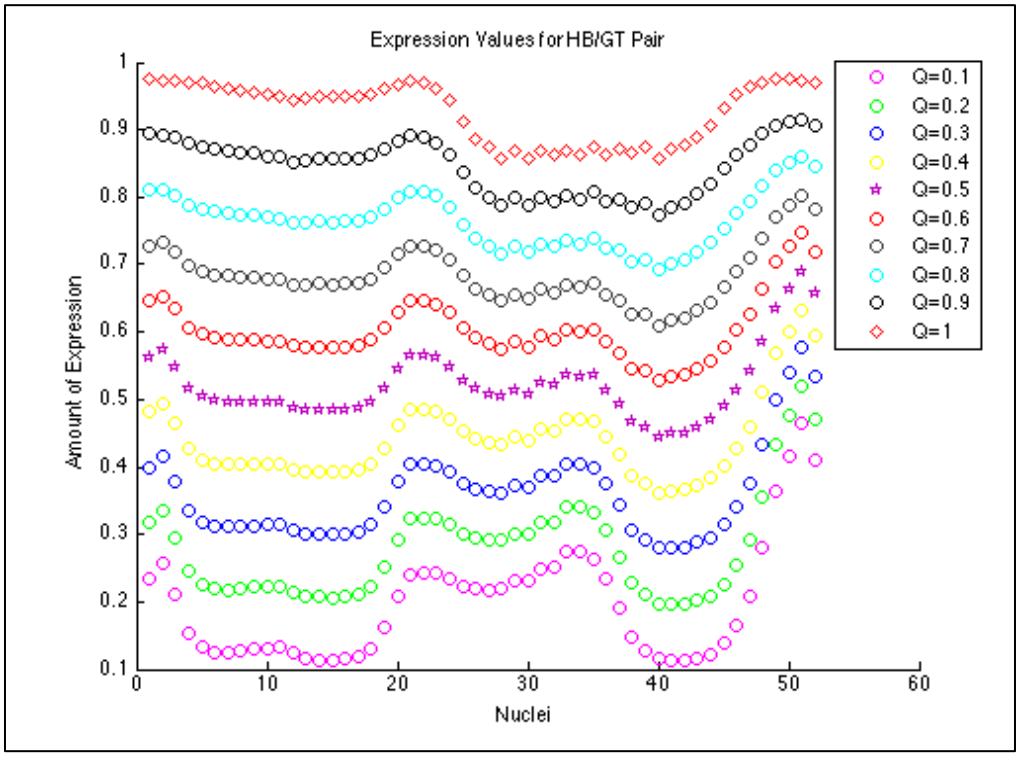


Figure 17. A predicted range of expression values an enhancer containing both a HB and GT binding site across the 52 nuclei of the developing embryo from anterior to posterior. The binding affinity constant for the activator is: 50. The binding affinity for the repressor is: 50. The quenching values (Q) range from 0.1-1. The y-axis of this graph represents the fractional probability of expression between 0 and 1. There are very low levels of predicted expression for this pairing. This makes sense because looking at the concentrations across the embryo in **Figure 6** they have relatively the same pattern of localization except for in the middle when HB has a greater concentration than GT.

CAD/GT Pair

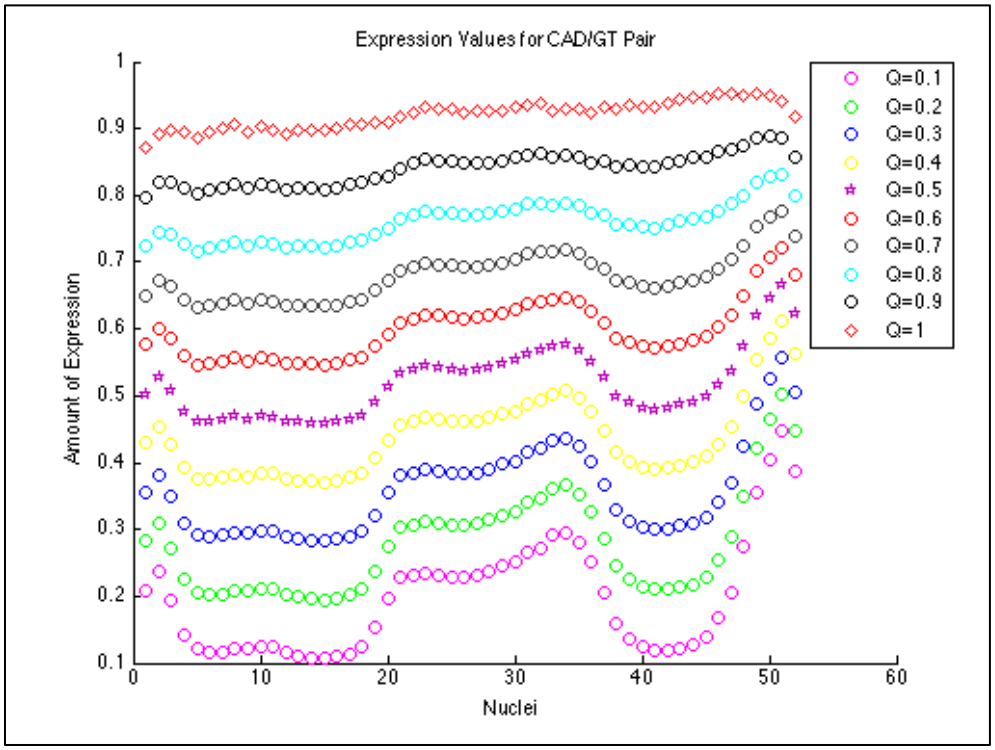


Figure 18. A predicted range of expression values for an enhancer containing both a CAD and GT binding site across the 52 nuclei of the developing embryo from anterior to posterior. The binding affinity constant for the activator is: 50. The binding affinity for the repressor is: 50. The quenching values (Q) range from 0.1-1. The y-axis of this graph represents the fractional probability of expression between 0 and 1. CAD is localized toward the posterior of the embryo. This is why the main expression we see is in the middle of the embryo and at the very posterior. GT has a strip of high concentration in the posterior of the embryo, which represents the dip and the steep increase at the most posterior.

BCD/KNI Pair

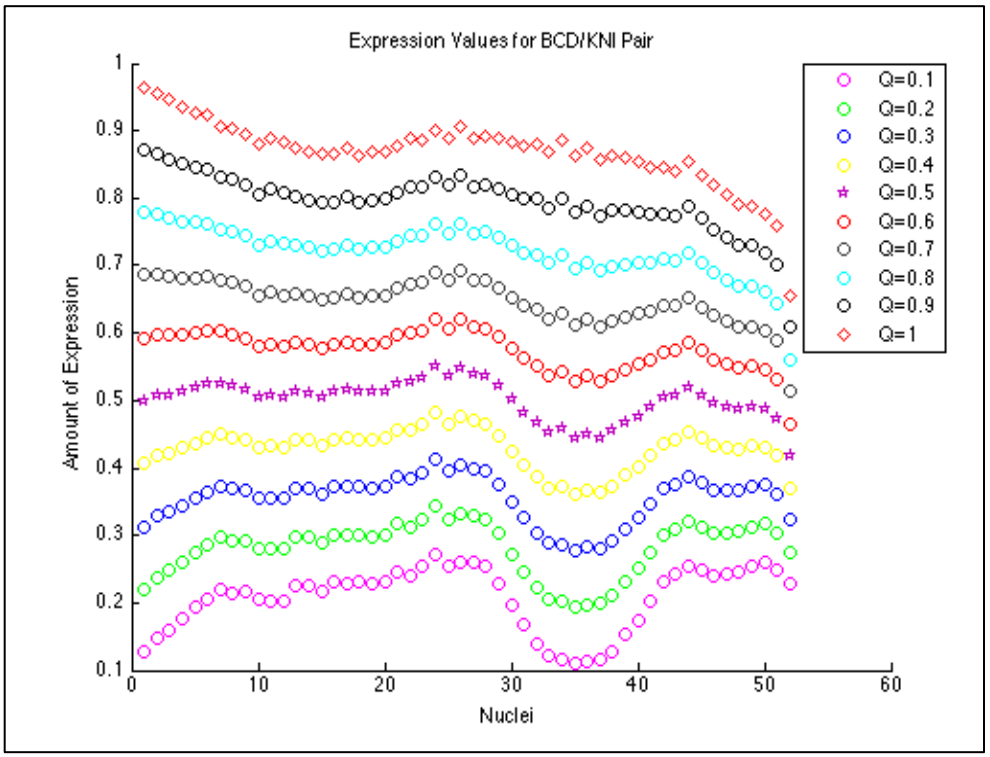


Figure 19. A predicted range of expression values for an enhancer containing both a BCD and KNI binding site across the 52 nuclei of the developing embryo from anterior to posterior. The binding affinity constant for the activator is: 50. The binding affinity for the repressor is: 50. The quenching values (Q) range from 0.1-1. The y-axis of this graph represents the fractional probability of expression between 0 and 1. There are low levels of predicted expression because there are really only low levels of BCD even in the anterior where it is located.

To obtain more reasonable binding affinity data for the single bound activator state, I utilized MATLAB again. After noticing that the values were much too high to represent real biological data, I decided to decrease the values that had been chosen. For each single bound activator state I decided to

run the model with K values between one and nine with increments of one. I ran this for each transcription factor: BCD, CAD and HB and the results are shown (along with a comparison of the activator concentrations) in **Figures 20, 21,** and **22**.

Lower K Values for a Single Activator State

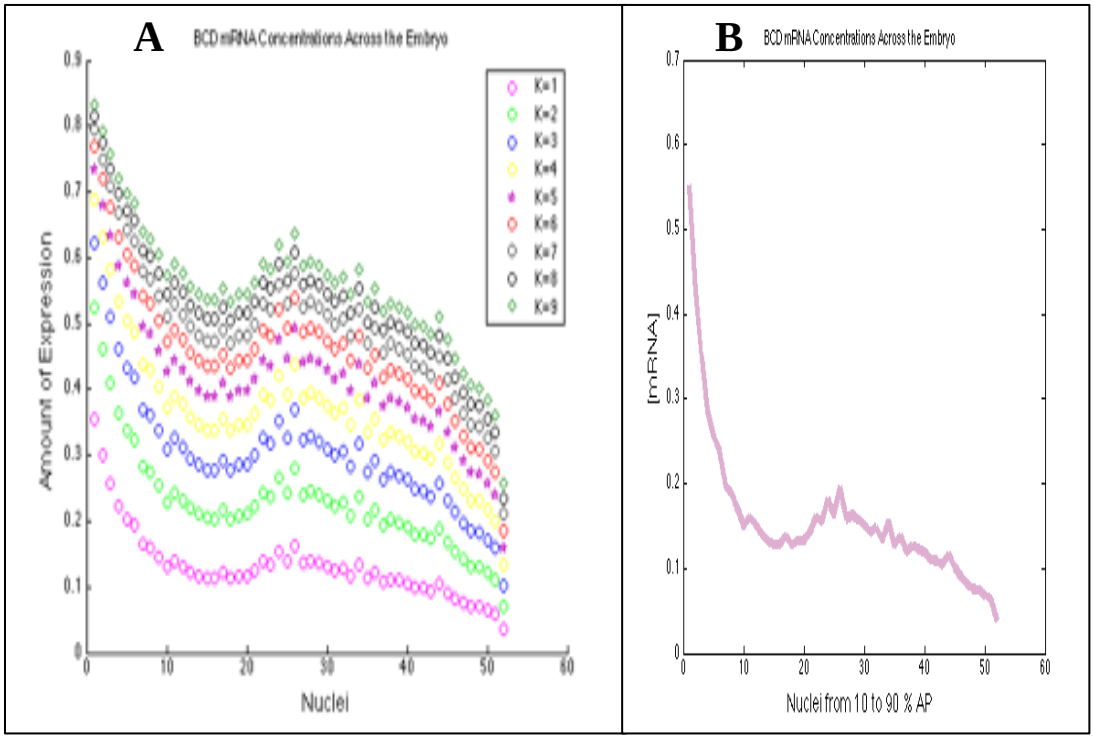


Figure 20. BCD with lower K values. Predicted expression values corresponding to an enhancer with a single BCD binding site across the embryo at time point two when the binding affinity is between 1 and 9 compared to the BCD transcription factor concentration gradient across the embryo. A) The mRNA predicted expression values corresponding to an enhancer with a single BCD binding site for the 52 nuclei across the embryo from anterior to posterior. B) BCD concentration gradient across the embryo. Comparing the two figures, they exhibit the same shape and the same pattern of expression for this single activator bound state.

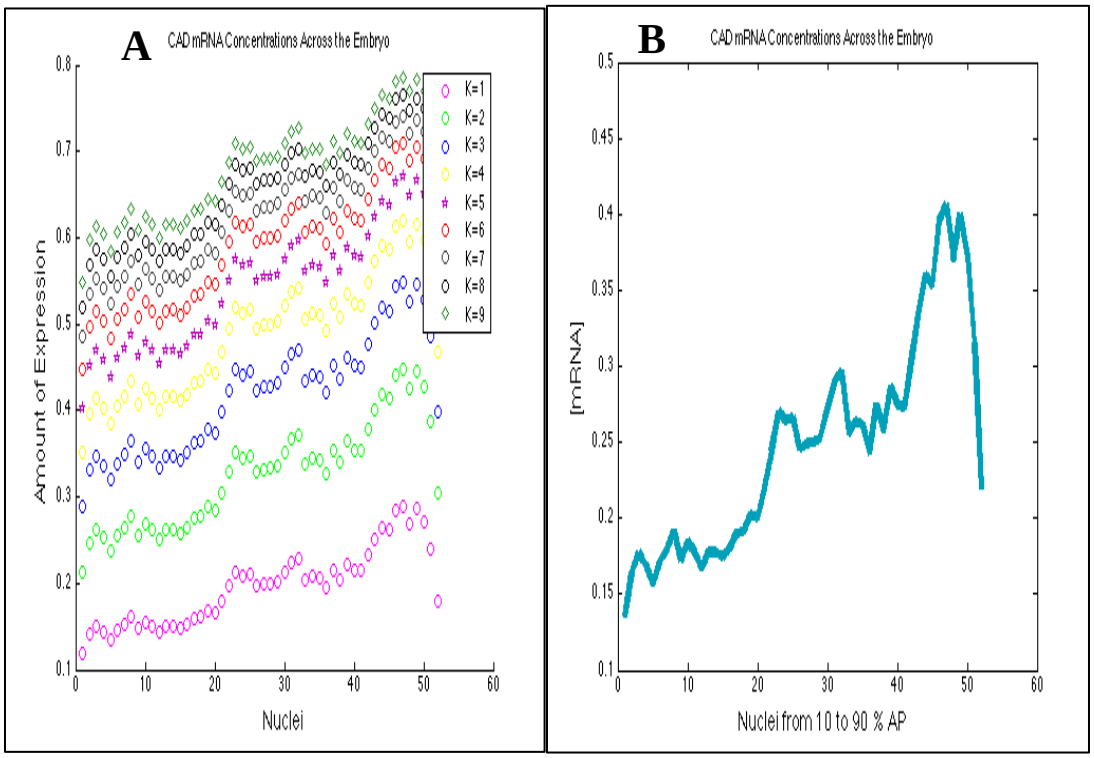


Figure 21. CAD with lower K values. Predicted expression values corresponding to an enhancer with a single CAD binding site across the embryo at time point two when the binding affinity is between 1 and 9 compared to the CAD transcription factor concentration gradient across the embryo. A) The CAD mRNA predicted expression values for the 52 nuclei across the embryo from anterior to posterior. B) CAD concentration gradient across the embryo. Comparing the two figures, they exhibit the same shape and the same pattern of expression for this single activator bound state.

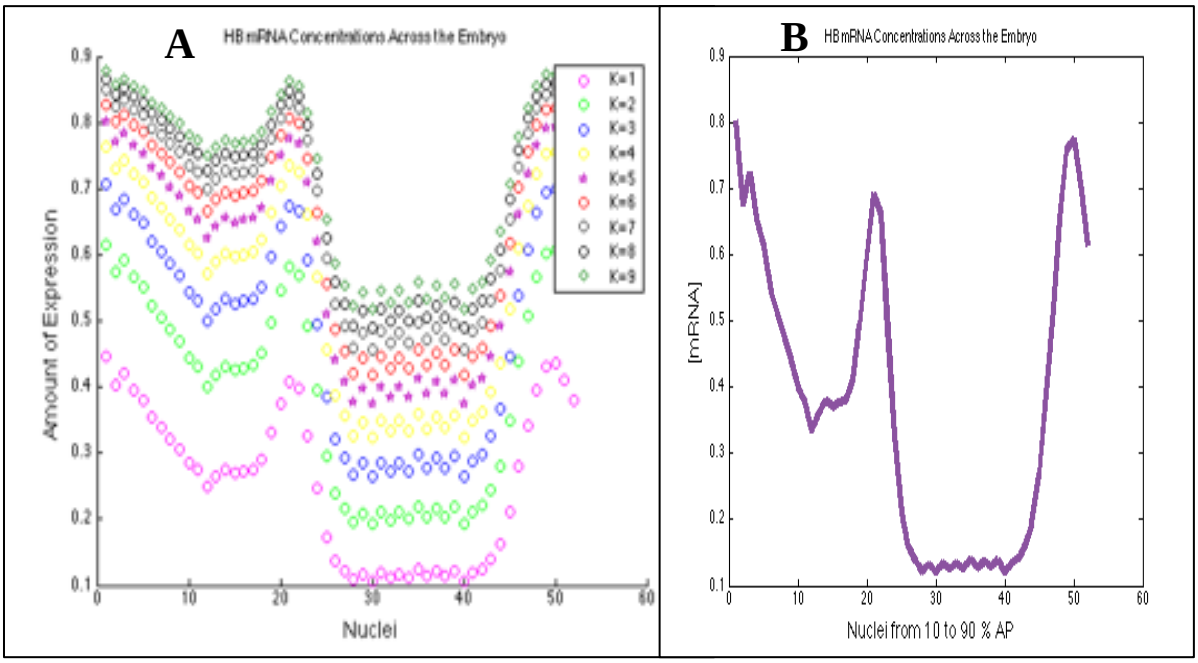


Figure 22. HB with lower K values. . Predicted expression values corresponding to an enhancer with a single HB binding site across the embryo at time point two when the binding affinity is between 1 and 9 compared to the HB transcription factor concentration gradient across the embryo. A) The HB mRNA predicted expression values for the 52 nuclei across the embryo from anterior to posterior. B) HB concentration gradient across the embryo. Comparing the two figures, they exhibit the same shape and similar patterns but the dissimilarity comes into play with the amount of expression between 30 and 40 nuclei in the embryo. The predicted expression does not drop as low as it does in the concentration gradient.

DISCUSSION

Although there is a lack of results from the biology portion of this experiment, the mathematical portion provides great insight to what the biological results might be expected to look like. These two results together would support the idea that quenching values and perhaps eventually cooperativity values are biologically relevant to transcriptional regulation and *Drosophila melanogaster* embryonic development.

The single activator code provides a glimpse at what might happen (both numerically and biologically) under the conditions of no repression. This gives a baseline binding affinity for certain transcription factors without involving other potential parameters. All of the results show that the higher the binding affinity, the higher the expression values which can be determined by the concentration of mRNA. This of course makes sense from the biological point of view because the more the activator is bound, the more mRNA will be made, and thus a higher level of expression.

Every transcription factor, activator and repressor alike, will have its own binding affinity, which we will assume is constant and unwavering. For most activators, I predict that the binding affinity value will remain quite low. This is due to the results seen in **Figures 20-22**. When the binding affinity values were low, I saw results that were similar to their localization patterns. Not only does this signify that this could be the binding affinity for that transcription factor, but we can also assume that because the shape and the

values are roughly the same, the mathematics behind the data are both biologically and mathematically sound.

It appears that for BCD and CAD, the binding affinity constant is roughly 2. This is not the case for HB. When comparing HB's concentration gradient to its predicted expression the values do not seem to line up as well as for the other two. Although the shape is the same, the level of expression never seems to drop as low as it should based on the biological data. After pondering this, I theorize that because HB is a concentration dependent activator or repressor its binding affinity along the nuclei of the embryo could potentially change as its concentrations fluctuate. Very low binding affinities (around 1) seem to allow for the lower levels of expression observed and a binding affinity of about 5 allows the higher level of expression seen with this transcription factor. We could also hypothesize that there is higher basal expression in the embryo, and we should try a higher value to represent the empty state. This of course should be further tested biologically.

Repressor binding affinity values are still undetermined based on these experiments, making it difficult to choose parameters for the quenching values. Despite the lack of knowledge about the binding affinity values of the repressors, similar patterns can be seen throughout all of the data. To reiterate, high quenching values make for poor repressors, and low values make for efficient repressors, therefore the mid-range quenching could be our target values for the future. Upper mid-range quenching values seem to signify the beginning of highly saturated expression and plateauing in the data. This

could mean that these biological quenching values may fall somewhere in the middle of the chosen range I chose to model. I would hypothesize that the quenching ability of the repressor would fall somewhere between the range of 0.2 and 0.6 before the saturation of expression is seen. This can be confirmed if the molecular biology portion of this experiment is carried out to the full extent.

If the binding affinity for an activator is too high, it may cause protein overexpression in the organism (as we see in the predicted expression from **Figures 12, 13** and **14**). Protein overexpression is a phenomenon where a single gene product is being synthesized in excessive amounts, which can cause problems later on in the organism's development (Brown, 2007). This is evident in the graphs because of the saturation of predicted expression values for extremely high K and Q values. This is another way to validate that our binding affinities should come out to be lower value and why it is also important to look into repression binding affinities, another way to decrease the risk of protein overexpression.

Running the codes with all of the parameters also yielded more satisfying results. When comparing the results from the graphs to the known concentrations of transcription factor mRNA, they seem to match up perfectly with the concentrations of activator in relation to its paired repressor concentrations. For example, looking at **Figure 16**, when the activator (BCD) was at higher levels of concentration than the repressor (KR), the levels of predicted expression were higher and vice versa. The low dip in the data is

seen when the concentration of KR was higher than for BCD, therefore repressing the (predicted) expression for *Eve*. For each graph I was able to saturate the amount of expression by increasing Q all the way to 1. This saturation was constant with high Q values across all of the graphs. With Q values approaching 1 it lost its ability to repress anything, and became an extremely poor repressor. This reveals that biologically, the Q values would be much lower. On the other hand, if the Q values were low, I saw a decent amount of repression in areas of high repressor mRNA concentrations. Again biologically, a low Q values means that it is a very good repressor, so logically this value should fall somewhere in the middle of these two extremes if we consider the idea that quenching is only slight amounts of repression. Another consideration that needs to be made is that this is only short-range repression because repressors have a preferred distance due to the known spatial constraints (Fakhouri *et al.*, 2010). This distance is about 100 base pairs long, so different ranges along this distance must be tested. Running these tests will have the capacity to show how an activator and a repressor truly work together in order to regulate transcription.

Future Work

Unfortunately, without the in-situ hybridization results there are no real biological data to compare these numbers to. In future-work, when someone has the ability to complete this experiment, they will be able to use these graphs as a comparison to their results. This comparison will lead to the eventual narrowing down of specific binding affinity values (for both the

activators and the repressors) as well as quenching values. Gaining this knowledge will lead to further insight on how quenching actually works, as well as its biological relevance in development.

More experimentation needs to be performed with MATLAB as well as carrying out the molecular biology. Although having already run lower binding affinity values, this number can be even more thoroughly pinpointed to an exact number. With the help of the molecular biological experiments and a solid pipeline for quantifying the amount of expression, these numbers can be determined as close to their true biological data as possible.

CONCLUSIONS

To determine the importance of this research we must take a closer look at the current body of knowledge surrounding quantitative data about binding affinities and quenching values. This knowledge is extremely minimal, as a bottom-up study is a new idea in this field of research. Previous studies that have been done can be compiled into the extensive background information in which a bottom-up study has the potential to gradually build upon. We know that the location of the repressor sites matter for the ability to quench or fully repress the transcription complex, and that when located within an upstream enhancer, a repressor has the ability to locally quench a nearby activator. This study was able to confirm the idea of short-range repression in gene regulation, as well as the idea spatial relationships between the CRMs is extremely crucial for proper control of transcription (Gray and Levine 1996).

More evidence for the idea of quenching through short-range repression is that knirps may be able to quench transcription, or locally inhibit upstream activators as one study shows in transgenic embryos (Arnosti *et al.* 1996). These studies about short-range repression show that not only having information about the locations of transcription factor binding sites (bioinformatically predicted) is important, but so is the ability of the activators and repressors to bind to these sites and regulate transcription. A mathematical perspective on this is crucial, as these numerical insights tell a tale of the true abilities. It is important to build off of this body of knowledge so that we can

learn more about the phenomenon of transcriptional regulation throughout development. Knowing the biomathematical binding affinity values may benefit our methods of studying conditions impacted by genetics if we know the particular transcription factors involved in the process.

To relate the idea of quenching and its importance in this field of study, researchers have been studying its effects in certain human diseases and viruses. One such virus is the human papillomavirus or HPV, which can sometimes lead to cancer. This virus' expression can be mediated by the transcription factor YY1, which can repress HPV transcription by quenching AP-1 activity. After much research, it was found that there are five sites required for the repression of HPV, and a mutation in any one of them will correlate to a six-fold increase in transcriptional activity (O'Connor *et al*. 1996). Knowing more about the binding affinity of YY1 and the actual quenching parameters can guide this study in a new and exciting way that could revolutionize the way we see and think about modern medicine.

In conclusion, thermodynamic-based models have the ability to provide a wealth of information about enhancer functional activities during embryogenesis in *Drosophila melanogaster*. Using bioinformatically obtained information about enhancer activities, as well as molecular biological techniques we can only hope to understand the very finite underpinnings of transcriptional regulation and the complexity of embryonic development. Gaining this knowledge will hopefully lead to advancements in many different scientific fields of study.

REFERENCES

Adkins, N., Hagerman, T., & Georgel, P. (2006). GAGA protein: A multi-faceted transcription factor. Biochemistry and Cell Biology, 84(4), 559-567.

Arnosti, D., Gray, S., Barolo, S., Zhou, J., & Levine, M. (1996). The gap protein knirps mediates both quenching and direct repression in the *Drosophila* embryo. The Embo Journal, 15(14), 3659–3666-3659–3666.

Brody, T. (2001). The interactive fly (27th ed.). Bethesda: Society for Developmental Biology.

Brown, A. (2013). A synthetic biology approach to test conserved regulatory motifs in *Drosophila melanogaster*. Harvey Mudd senior thesis publication.

Brown, T. (2007). Genomes 3. New York: Bios Scientific.

Borok, M., Tran, D., Ho, M., & Drewell, R. (2010). Dissecting the regulatory switches of development: Lessons from enhancer evolution in *Drosophila*. Development, 137, 5-13.

Dresch, J., Drewell, R. (2012). Decoding the cis-regulatory grammar behind enhancer architecture. Unpublished book chapter.

Fakhouri, W., Ay, A., Sayal, R., Dresch, J., Dayringer, E., & Arnosti, D. (2010). Deciphering a transcriptional regulatory code: Modeling short-range repression in the *Drosophila embryo*. Molecular Systems Biology, 6(1), n/a.

Gilbert, S. (2000). The genetics of axis specification in *Drosophila*. In Developmental Biology (6th ed.). Sunderland, MA: Sinauer Associates.

Gray, S., & Levine, M. (1996). Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in *Drosophila*. GENES & DEVELOPMENT, 10, 700-710.

Griffiths, A., Wessler, S., Carroll, S., & Doebley, J. (2012). Introduction to genetic analysis (10th ed., p. 9, 435). New York, New York: W.H. Freeman and Co.

Helden, J. (2003). Regulatory Sequence Analysis Tools. Nucleic Acids Research, 31(13), 3593-3596.

Kim, S., Shim, H., Jeon, B., Choi, W., Hur, M., Girton, J., ... Jeon, S. (2011). The pleiohomeotic functions as a negative regulator of *Drosophila* even-skipped gene during embryogenesis. Molecules and Cells, 32(6), 549-554.

Kim, Y., & Lis, J. (2005). Interactions between subunits of *Drosophila* Mediator and activator proteins. Trends in Biochemical Sciences, 245-249.

Laity, J., Lee, B., & Wright, P. (2001). Zinc finger proteins: New insights into structural and functional diversity. Current Opinion in Structural Biology, 11(1), 39-46.

Li, L., & Arnosti, D. (2011). Long- and Short-Range Transcriptional Repressors Induce Distinct Chromatin States on Repressed Genes. Current Biology, 21(5), 406-412.

Ludwig, M., & Kreitman, M. (1998). Evolutionary dynamics of the enhancer region of even-skipped in *Drosophila*. Molecular Biology and Evolution, 12(6), 1002-1011.

Luengo, C., Keränen, S., Fowlkes, C., Simirenko, L., Weber, G., DePace, A., . . . Knowles, D. (2006). Three-dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution I: Data acquisition pipeline. Genome Biology, 7.

O'Connor, M., Tan, S., Tan, C., & Bernard, H. (1996). YY1 represses human papillomavirus type 16 transcription by quenching AP-1 activity. Journal of Virology, 70(10).

Passarge, E., Horsthemke, B., & Farber, R. (1999). Incorrect use of the term synteny. Nature Genetics, 23, 387-388.

Potter, H., & R. Heller. (2010). Transfection by Electroporation. Current Protocol Molecular Biology. Chapter 9.3.

Ptashne, M. (1986). Gene regulation by proteins acting nearby and at a distance. Nature, 322(6081), 697-701.

Roberts, D. (2006). *Drosophila melanogaster*: The Model Organism. Entomologia Experimentalis Et Applicata, 121(2), 93-103.

Sanson, B. (2001). Generating patterns from fields of cells: Examples from *Drosophila* segmentation. EMBO Reports, 2(12), 1083-1088.

Sherman, M., & Cohen, B. (2012). Thermodynamic State Ensemble Models of cis-Regulation. PLoS Computational Biology, 8(3), E1002407-E1002407.

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

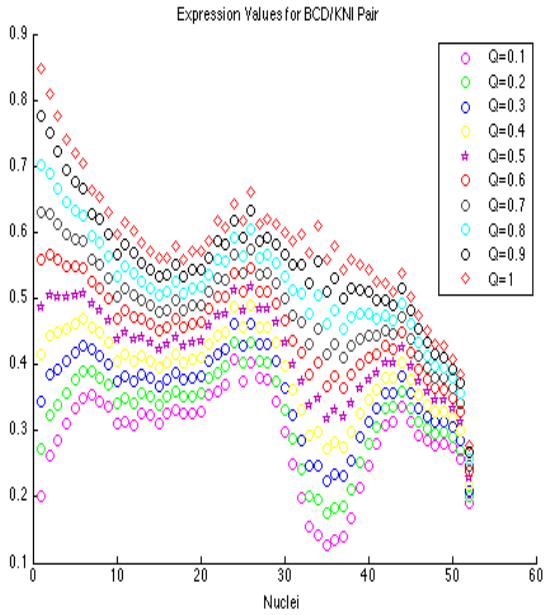
Venken, K., & Bellen, H. (2005). Emerging technologies for gene manipulation in *Drosophila melanogaster*. Nature Reviews Genetics, 6, 167-178.

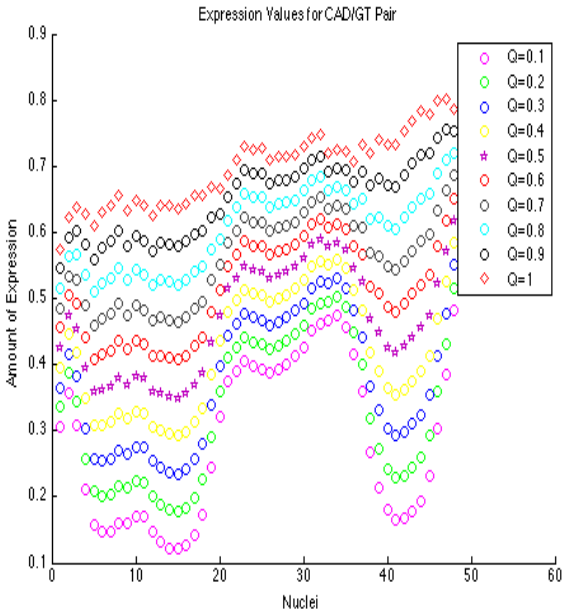
Wolffe, A., Wong, J., & Pruss, D. (1997). Activators and repressors: Making use of chromatin to regulate transcription. Genes to Cells, 2(5), 291-302.

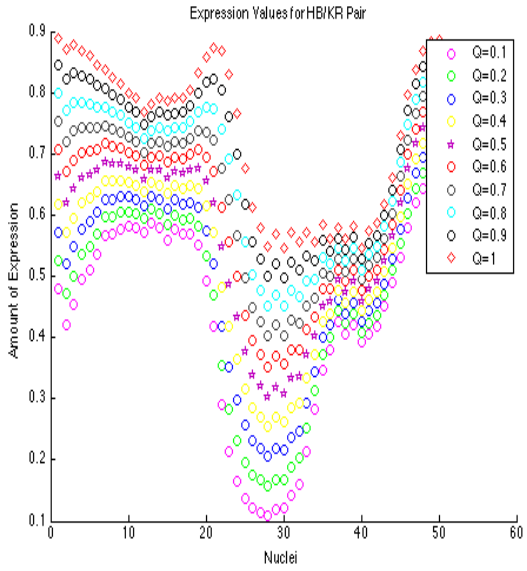
Yavatkar, A., Lin, Y., Ross, J., Fann, Y., Brody, T., & Odenwald, W. (2008). Rapid detection and curation of conserved DNA via enhanced-BLAT and EvoPrinterHD analysis. BMC Genomics, 9, 106.

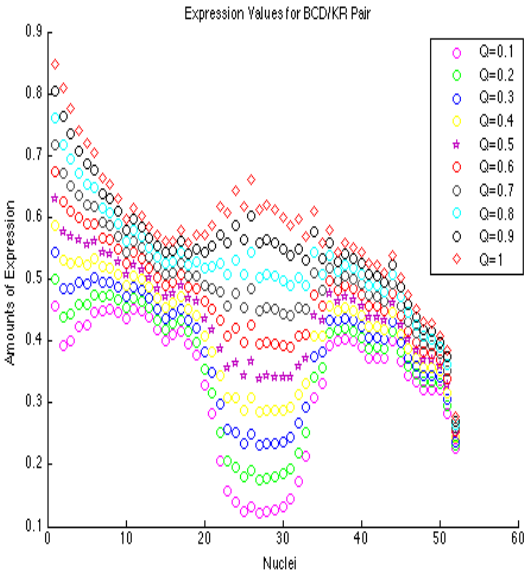
APPENDIX

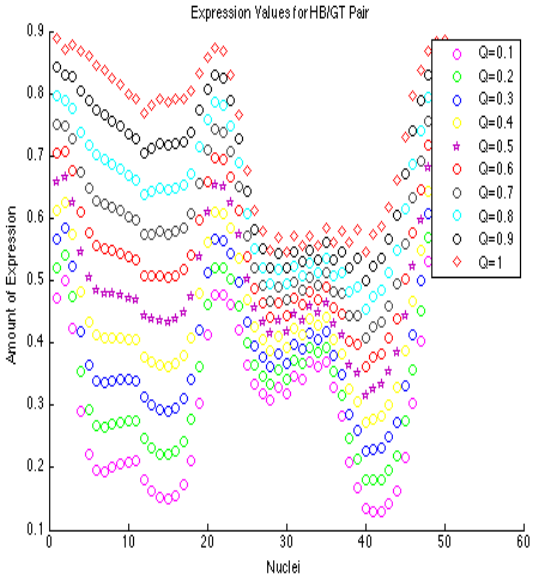
KA = 10 KR=10 for all graphs










Expression Values for HB/GT Pair
Amount of Expression
Nuclei
Q=0.1
Q=0.2
Q=0.3
Q=0.4
Q=0.5
Q=0.6
Q=0.7
Q=0.8
Q=0.9
Q=1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
10
20
30
40
50
60