Relating the Expression-based and Sequence-based Estimates of Regulation in the Gap Gene System of *Drosophila melanogaster*

Faiyaz Al Zamal

Master of Science

School of Computer Science

McGill University

Montreal,Quebec

2007-06-26

A thesis submitted to Mcgill University in partial fulfillment of the requirements of the degree of Master of Science

Your file Votre référence
ISBN: 978-0-494-51358-3
Our file Notre référence
ISBN: 978-0-494-51358-3

DEDICATION

This document is dedicated to my fiancée Noor-E-Tamannah.

ACKNOWLEDGEMENTS

This thesis is a product of a year of a pleasant and productive collaboration with my supervisors, Professor Theodore J. Perkins and Professor Mathieu Blanchette. First and foremost, I thank both of them for the wonderful support they have given me throughout the year. Their encouragement and enthusiasm was the driving force behind the thesis. Their suggestions and advices aided me at times when I got stuck while solving complex problem. They have been very generous to provide me with the funds, without which it would not have been possible for me to continue my research.

This work was supported in part by grants from NSERC Canada to Professors Theodore J. Perkins and Mathieu Blanchette. I take the opportunity to thank NSERC Canada for their support in Natural Sciences and Engineering research.

I am grateful to Professor Saurabh Sinha of UIUC who was generous enough to provide me with his code of the latest version of Stubb that we have used extensively in Chapter 5. I also thank Muhammad Jamal Sheikh for his help for the pre-processing step of the linear regression results in Chapter 3. I must mention the contributions of the members of Mcgill Centre of Bioinfomatics who were always cooperative and supportive towards me. Especially I would like to thank the members of Mathieu Blanchette's Lab. The weekly meetings gave me opportunities to present my work and more importantly enhanced my breadth and depth of knowledge on various branches of Bioinformatics. I also thank Gerry Chevalier for her ever smiling and supportive attitude.

I am thankful to all the Bangladeshi students of Mcgill. Especially I am grateful to Shah Asaduzzaman, Junaed Sattar and Sadaf Mustafiz. Their help in academic and non-academic affairs in the last two years aided me at times of difficulties.

Finally I thank my parents who funded me for my first year in Mcgill. Without their support, I would not have got an opportunity to pursue my studies at Mcgill. The wonderful inspiration that I got from my parents, my fiancée and my other family members deserves a very special thanks on my part.

ABSTRACT

Quantitative analysis of *Drosophila melanogaster* gap gene expression data reveals valuable information about the nature and strengths of interactions in the gap gene network. We first explore different models for fitting the spatiotemporal gene expression data of *Drosophila* gap gene system and validate our results by computational analysis and comparison with the existing literature. A fundamental problem in systems biology is to associate these results with the inherent cause of gene regulation, namely the binding of the transcription factors (TF) to their respective binding sites. In order to relate these expression-based estimates of gap gene regulation with the sequence-based information of TF binding site composition, we also explore two related problems of i) finding a set of regulatory weights that is proportional to the binding site occupancy matrix of the transcription factors in current literature and ii) finding a set of position weight matrices of the TFs that produce a new binding site occupancy matrix showing a greater level of proportionality with our regulatory weights. Our solution to the first problem yielded a regulatory weight matrix incapable of explaining the true causes of gene expression profile despite its relative numerical accuracy in predicting the gene expressions. On the other hand, the second optimization problem could be solved up to a reasonable level of accuracy, but further analysis on the result demonstrated that this optimization problem may be under-constrained. We devise a simple regularization strategy that helps us to reduce the under-constrained nature of the problem.

ABRÉGÉ

L'analyse quantitative du niveau d'expression des gènes du système gap de *Drosophila melanogaster* révèle d'importantes informations quant 'a la nature et la force des interactions entre les membres du réseau. Nous explorons d'abord diffrents modèles pour l'ajustement de données d'expression spatiotemporelles du système gap de la drosophile et validons nos résultats grâce à une analyse bioinformatique et basée sur la littérature. Un problme fondamental de la biologie des systèmes est l'association de ces résultats à leurs causes inhérentes, soit la liaison de facteurs de transcription à leurs sites respectifs. De manière relier les niveaux d'expression observés des gènes du système gap à la composition de leur séquences régulatrices, nous explorons deux problmes: i) la recherche de poids de rgulation qui soient proportionels aux prédictions de sites de facteurs de transcription trouvés dans la littérature, et ii) la recherche de nouvelles matrices de poids de facteurs de transcription qui résultent en des prédictions de sites qui démontrent un haut niveau de proportionalité avec nos poids de régulation. Notre solution au premier problème donne une matrices de poids de régulation qui est incapable d'expliquer la cause réelle du profile d'expression observé, malgré une bonne précision des niveaux d'expression. Par ailleurs, le second problème d'optimization peut être résolu jusqu'à un niveau de précision acceptable, mais l'analyse des résultats démontre que le problème est sous-contraint. Nous avons créé un algorithme de régularisation qui aide à réduire la sous-détermination du problème.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Biological Overview

1.1.1 Gene Regulation Mechanism

The genome of a living organism carries the whole repertoire of information essential for controlling all the cellular processes. This information is encoded in the genome DNA sequence. From a Computer Scientist's viewpoint, the genome of any organism can be thought of as a very long *string* whose characters are taken from an alphabet consisting of four characters A, C, G and T which represent the *nucleotides* Adenine, Cytosine, Guanine and Thymine respectively. This long DNA sequence can be broadly categorized into *coding region* and *non-coding region.* Both the *coding* and *non-coding regions* consist of shorter fragments of non-overlapping sequences. The *coding region* of a genome encodes numerous functional products, most commonly the necessary proteins for all the biological processes. The *non-coding region* of the genome contains the essential condition and information required to control the production of genomic products. A gene is a segment of DNA sequence that usually produce a single protein, although a single gene can produce multitude of proteins. The prokaryotic genes are continuous strings, but eukaryotic genes usually have coding *exons* separated by several non-coding *introns.*

The number of genes in the genomic DNA sequence varies greatly across different species. Yeast has only 6000 genes, human or mouse genome has about 25,000 while

Poplar genome has about 300,000 genes. However, all the genes are not always transformed into the corresponding protein. When a gene is transformed into the gene product(protein) that it codes, it is said to be *expressed* and the corresponding process is called *gene expression.* The non-coding region of the genome encrypts the information about *gene regulation*, a process which determines which gene is to be expressed (or not expressed) under the presence or absence of a certain biological condition, the timing and amount of the protein production. For a particular gene, the sequence of nucleotides that encodes its gene regulation information is called the regulatory region of that gene. The regulatory region is usually located at a close proximity of the target gene.

The process of gene expression and regulation is a complex one, although a simpler high level view can be presented. In case of prokaryotes, the first step is *transcription* where the DNA sequence is *transcribed* into messenger-RNA (mRNA) with the help of an enzyme called RNA-polymerase which attaches itself to a short sequence in the promoter. This mRNA is then *translated* into protein sequence by the *translation* process. For eukaryotes, the transcription process results in pre-mRNA which are *spliced* to obtain mRNA such that only the exons contribute towards the final protein product and the intronic pre-mRNA are spliced out. The translation process then translates the mRNA into the final protein. Proteins are sequences of amino acids. There are 20 amino acids, therefore a protein can be viewed as a string whose characters are drawn from an alphabet of 20 characters. As translation involves a transformation from a 4 character alphabet to a 20 character alphabet, three consecutive nucleotides (also known as a *codon*) code a single amino acid.

The gene regulation process starts when the RNA polymerase attaches itself to the promoter. The attachment of RNA polymerase to the regulatory region is affected by the presence of certain proteins that are produced by other genes. These proteins may attract the RNA polymerase and act as a catalyst for the regulation process or they may cause a hindrance in the process and act as a repressor. The proteins acting either as activator or repressor are called the transcription factors (TF) for the particular gene. A TF bind to a set of specific sequences of nucleotides, usually 5-15 base pair in length, present in the regulatory region of the genes. These sequences are called the transcription factor binding sites (TFBS). The set of TFBS attracting a specific TF is often referred as a *regulatory motif* or *binding site motif* in the literature, as it has been observed that different sequences attracting a particular TF are always of the same length and the nucleotides at different position of the sequences often match. Therefore the gene regulation process depends on the presence (or absence) of the binding site motifs and the availability(or non-availability) of the transcription factors for the gene.

In the case of most eukaryotic genes, the expression pattern of a particular gene is controlled by several cis-regulatory modules (CRMs), each of which consists of multiple TFBSs of usually more than one TFs [68]. Each CRM acts independently to drive the expression profile of its targets at a specific position on the body of the organism. In our work we computationally analyze the *Drosophila melanogaster* gap gene regulatory network expression data and the genome sequence data of the *Drosophila* using several models and hypothesis and aim at ascertaining the precise relationship among the expression pattern of a given gene, the expression pattern of all the other genes who act as TFs for the first gene and known binding site composition of the known CRMs.

1.1.2 *Drosophila* Segmentation Gene System

The segmentation gene network of *Drosophila Melanogaster* is a widely used paradigm for the analysis of gene expression and transcriptional control in eukaryotic cells. The segmentation genes are the genes responsible for the formation of the segmented body pattern of *Drosophila.* (For a review please refer to [1, 28]). This body pattern is determined during the syncytial blastoderm stage, a stage characterized by the presence of multiple nuclei inside a single cell, of *Drosophila* embryo [57]. The segmentation genes are classified into several types. The dependency among different types of segmentation genes work in a well defined hierarchical manner which is described in the next paragraph. Figure 1-1 taken from Schroeder et al. [55] schematically represents the hierarchy of dependencies among several different gene types of the segmentation gene network.

The *Drosophila* egg, once fertilized, forms the larva within 24 hours. During the early stage of the development(by 3 hr), the zygotic nuclei divides rapidly but the cellular membranes are yet to be formed [60]. This state is called the syncytial blastoderm stage of the embryo. During this stage, most nuclei transport to the surface of the eggs and become active participator in the process of transcription. The maternal input factors, synthesized during oogenesis, act as morphogens (primary stimuli for expression pattern formation) in this state and gives rise to the early activation of *gap gene* expression. The maternal factors are spatially distributed as a slowly increasing/ decreasing gradient along the anterior posterior (ap axis). The

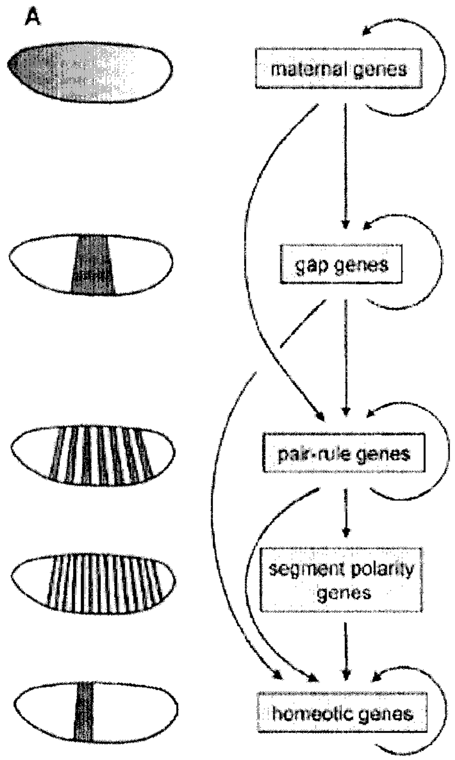


Figure 1–1: The hierarchy of interactions in the Segmentation genes of *Drosophila melanogaster* [55].

zygotic *gap genes* are a set of segmentation genes which are expressed along one or more broad and overlapping domain. They affect several contiguous segments [53]. The maternal and gap genes together drives the expression of pair-rule genes that are expressed in seven complete alternative segments. The segment polarity genes, divided into 14 stripes, represent the final segments and are simultaneously controlled by *gap genes* and *pair-rule genes.* The homeotic genes are responsible for identification of the segments.

The *Drosophila* gap gene system is the particular system that we analyze in our study. The maternal morphogen Bicoid (Bcd), Caudal (Cad) and Hunchback (Hb) stimulate the zygotic expression of the gap gene Hunchback (*hb*), Kruppel (*kr*), Giant(*gt*) and Knirps (*kni*). The protein product of the terminal gap gene Tailless (Tll) also acts as a TF for these gap genes. Moreover, gap genes regulate each other and most of these gap gene to gap gene interactions are repressive. There is evidence of auto-regulation among the gap genes [42, 29, 30, 53, 50]. Due to the active interaction between these TFs, the gap genes attain their precise set of spatial domain of expression along the ap axis during cleavage cycle 14A.

At this stage, the cell membranes begin to form, but the nuclei are not yet completely surrounded by membrane. This is the cellular blastoderm stage, when the pair rule genes continues evolving into a well defined pattern of expression. However, the pair rule gene pattern does not culminate into its final pattern during the cellular blastoderm stage. During the gastrulation stage, the cell membranes completely surround the nuclei and the formation of the pair rule gene pattern continues. During the germ band extension stage, the pair rule gene products slowly decays, but segment polarity genes are expressed which, together with pair-rule genes, regulate the homeotic gene expression. Thus the final segmented body pattern of *Drosophila* is established.

In the next section we briefly discuss some selected works on modelling gap gene network and computational approaches towards detecting TFBS.

1.2 Previous Work

1.2.1 Modelling the Gap Gene Regulatory Network

The motivation behind modelling gene regulatory network is to build a thorough understanding of the gene regulatory mechanism governing an important biological process. The gap gene regulatory network plays a crucial role in the formation of the segmented body pattern of *Drosophila.* Moreover, the gap gene system is the first set of segmentation genes that directly responds to maternal morphogens and the pattern formation is almost entirely due to transcriptional regulation [42, 55]. Therefore, this network has been studied extensively in the last couple of decades. These studies can be broadly categorized as qualitative and quantitative analysis of the regulatory effects of various transcription factors on different target genes. Most of these studies concentrated on the transcriptional regulation for either the anterior-posterior-axis (*ap*-axis) and or the dorsal-ventral-axis (*dv*- axis) pattern formation as these two systems are considered independent of each other.

a **Qualitative Approaches** The early studies of the gap gene regulatory network were principally qualitative. Qualitative modelling of the regulatory system is much simpler than a quantitative modelling and so is its analysis and interpretation. Moreover, the dearth of exact quantitative experimental data of various gap gene expression under different condition motivated these qualitative studies. [50, 53]. Generally, the mode (activation/ repression) of the regulatory effect of a TF is assumed by

i analyzing the difference between mutant and wild-type expression patterns., and ii analyzing the transcription factor binding site composition.

For general GRNs, different qualitative approaches have been proposed and applied. The *boolean network* [64] , *stochastic logical networks* [67], *generalized logical formalisms* [50, 53] etc. are examples of such different models. In our study, we often compare our results with those of Rivera-Pomar et al. [50] and Sanchez Thieffry et al. [53]. In Rivera-Pomar et al., the authors descriptively explain the regulation of different genes by analyzing the mutant phenotypes of the embryo under different conditions and the DNA binding motifs present in the regulatory regions of those genes. Sanchez et al. [54, 53] is motivated by the work of Kauffman et al. [33] and Thomas et al. [64] which introduced a boolean formalization of the general gene regulatory networks. It presented the gene expression profile using logical discrete random variables where different thresholds of expression values represented different values for the discrete variable, i.e. 0 to represent no expression, 1 for low level of expression, 2 for high level of expression etc. At first, the authors construct a interaction graph among different TFs and their targets by studying the results of different mutant expression patterns. Then they formalize a system of generalized logical equations for the interaction graph. The embryo is divided into four non-overlapping regions A to D where region A corresponds to the anterior most portion and the region D refers to the posterior most portion. A state table can be constructed on the basis of these logical equations. This state table identifies all the stable states within the system and helps simulating the qualitative effect of loss-of-function and *cis*-regulatory mutations.

The qualitative studies mentioned above have been successful in identifying many interactions within the network with a relative degree of accuracy. However, they have often lead to ambiguity while predicting the interactions. This ambiguity is caused by the fact that the gap genes regulate each other, so the target genes are TFs themselves. Therefore, any observed effect can be explained by either a direct effect of a TF on the target or an indirect effect through a series of TF-target gene interactions. This ambiguity inspired researchers to employ quantitative approaches to model gene regulation with the required level of precision and determinism. The growing availability of the quantitative gene expression data bolstered their efforts.

b **Quantitative Approaches** As the transcriptional regulation process in an eukaryote is substantially complex and involves many layers of interactions, it is not possible to abstract it perfectly into a model. The choice of a particular model to represent a complex process depends largely upon the amount and nature of experimental data available and the objective of the assay [8]. GRNs have been modeled using stochastic equations [2, 19, 38, 49], Ordinary/ Piecewise Linear/ Partial Differential Equations. However, the most common and successful model used for gap GRN is *Gene Circuit Model* [39, 48, 24, 23] where the transcription rate change of a given gene is modelled by a combination of production, decay and diffusion of the gene product.In these models, the effects of all the TFs on every target gene is represented by a regulatory weight matrix. The signs of the weights denote the nature of transcription regulation, i.e activation or repression, while the magnitudes of the weights represent the strength of the regulatory influence. In our study, we look closely on the results of two such studies, one by Perkins et al. [42] and the other by Jaeger et al. [29, 30] . Both of these works analyze only the wild type expression profile (details in Chapter 2) and both of them uses differential equation models that deterministically represent the change in protein concentration as a combined result of production, decay and diffusion of the transcription factors. These parameters to define production, decay and diffusion are unknown, so the authors employ different optimization procedures for searching a suitable set of parameter values to fit the model. The aim of the optimization procedure is to find out a unique set of parameter values such that a simulated expression profile of the target gap genes matches as closely as possible to the observed profile of the genes, i.e. the sum of the squared error between the model prediction and actual observation is minimized. Jaeger et al. [29] uses Parallel Lam Simulated Annealing (PLSA). The algorithm is described in [48, 11] . Perkins et al. [42] concocts a three-step-strategy for the same purpose. The strategy adopted in Perkins et al. requires much less time for optimizing the fits. Qualitatively speaking, both these models are mostly successful in capturing the dominant interactions in the *Drosophila* gap gene network although the time required for fitting the model is a concern for Jaeger et al. models. However, the main drawback of such model is the use of phenomenological 'regulatory weights' with no connection to the underlying molecular process governing the binding of the TFs to the regulatory region sequence. In the subsequent chapter, we discuss more about the results of Perkins et al. and Jaeger et al. when we compare their results with various results that we get. Table 1-1 compares some different features of some selected qualitative and quantitative approaches.

Table 1–1: Comparison between different qualitative and quantitative models

Model	**Data used for analysis**	**Considers promoter (regulatory region) sequence**	**Model type**
RPJ [50]	Qualitative wild-type+mutant expression	Yes	Logical-static
ST [53]	Qualitative wild-type+mutant expression	No	Logical-dynamical
Jaeger et al. [29]	Quantitative wild-type expression	No	Nonlinear ODE/PDE
Perkins et al. [42]	Quantitative wild-type expression	No	Nonlinear ODE/PDE

1.2.2 Detection of TFBS

Transcription factors recognize specific regulatory motifs in the regulatory region of their targets. Scientists have identified many such TFBS sequences for various TF at the regulatory region of different target genes in the genome. These regulatory motifs are usually defined by means of Position Weight Matrix(PWM) or Position Specific Scoring Matrix (PSSM) [22] and consensus sequences. Position weight matrices list the expected frequency of the occurence of different nucleotides at different positions of the binding site motif. There are different approaches for finding the regulatory motifs from the genome, such as greedy algorithms (CONSENSUS [18]), Expectation Maximization (MEME [3]) and Gibbs Sampling method (GibbsDNA [37]). However, significant difficulties arise while detecting such short and degenerate sequences from the long regulatory regions. There are indeed a lot of *hits*, but further analysis revealed that only a small portion of them represent authentic binding sites. To circumvent these difficulties, many different approaches have been proposed and tested. Many of these methods attempt to use the expression data ([9, 56, 26, 66, 37, 16, 51, 46, 6, 5]). Some of these methods cluster the genes according to their expression profile and then concentrate on finding the presence of shared sequence motifs in the regulatory region of the genes belonging to the same cluster. Some other methods calculate the correlation of the expression profile of a gene with the occurrence of different small sequences in its regulatory region and identifies the small sequences showing significant correlation with the expression as binding sites for the gene. There are some other efforts on analytically modelling the underlying chemical process of binding instead of attempting to find the individual binding sites. However for eukaryotic genomes, most recent approaches focus more on finding CRMs instead of individual TFs [13] . A CRM usually consists of multiple binding sites for different transcription factors clustered into relatively small length sequence(<1000 bp). Searching for CRMs instead of individual binding sites reduces the likelihood of detecting false positive binding sites [55, 7].

1.3 Thesis Objective and Organization

We focus our concentration on two related problems. The first problem is to computationally identify the regulatory interactions and the extent of their effects in determining the key features of the gene expression profile. The second problem is to associate these regulatory interactions with the primary cause of gene regulation, namely, the binding of the transcription factor to their respective binding sites located in the regulatory region of the target gene.

In Chapter 2, we provide a brief introduction to the data used for our analysis. We also provide an overview of the general methods that we use to solve the above mentioned problems. In Chapters 3 and 4, we address the first problem of identifying the regulatory relationships. Chapter 3 deals with static models that analyzes only the final time point data. Chapter 4 discusses dynamic models that consider the expression data at all the time points. Chapter 5 describes our efforts to solve the second problem of associating the regulatory weights TF binding. Finally in Chapter 6, we conclude our thesis and discuss about possible directions for future research on this subject.

CHAPTER 2
Data & Methods

2.1 Data

2.1.1 Gene Expression Data

The quantitative gene expression data that we have used is available online in the FlyEx database [43]. This database contains the fluorescence labelled images of the both wild-type and mutant *Drosophila melanogaster* embryo at different time and space in cellular resolution. We use the quantitative wild type expression profile of the protein products of several different genes in the early *Drosophila* embryo during the late syncytial blastoderm stage of its development. The expression profile data is acquired by means of processing of the blastoderm stage embryo images that are obtained by immunofluorescent staining and confocal microscopy procedure. Several steps of acquisition and processing are performed before reporting the expression profiles. The steps are briefly described below.

i **Antibody Staining & Confocal Microscopy** : The wild type *Drosophila* embryos are immunostained with different fluorescence-tagged antibodies for three different gene products at a time. After that, the embryo confocal images are taken using a laser confocal scanning microscope. Images are taken using four different channels of the confocal microscope, and for each channel, two raw images are taken corresponding to two optical sections of the embryo. The gain is adjusted such that for each gene product, the maximum intensity were at 255 on a 8 bit scale [35].

ii **Image Segmentation** : The fluorescence intensity level of each nucleus is extracted by edge detection and the use of watershed operation for error correction [40].

iii **Background Removal** : The quantitative data is then normalized such that the distortions caused by the background signal are minimized. The background is approximated using a two dimensional paraboloid which is fit to the data [32].

iv **Temporal Classification** : The data is classified based on the age of the embryo. The cleavage cycle 9 to 13 are short spanned. However, the cleavage cycle 14A is significantly longer. Therefore the cleavage cycle 14A is divided into 8 temporal equivalent classes where each class (apart from the first two) represents approximately 6 minutes [41].

v **Registration** : Registration is the process of transforming the coordinates of different embryo images such that the characteristics features of each gene products are superimposed. Two independent methods of registration are used, i) Quadratic splines and ii) Wavelet transform [32].

vi **Averaging** : The variability of the data is mostly on the A-P axis. That is why the data can be represented as a 1D reference at cellular resolution. For each A-P axis position, the representative value of each gene expression is computed by averaging the expression profiles at the central 10% strip of the embryo along the D-V axis.

We have used the same dataset as used in [42, 29] .The data comprise of quantitative wild type concentration profile for the protein products of *bcd, cad, hb, Kr, gt, kni* and *tll.* Only the cleavage cycle 13 and 14A data is used, which covers a 68 minute time span starting from the first detection of gap proteins until the onset of gastrulation [29]. Hunchback expression data for cleavage cycle 12 is used as the *hb* initial expression level ($t = 0$). The other gap genes are not expressed at all before cleavage cycle 13. We only consider the positions along the trunk region of the embryo (35% to 92% of the embryo length, total 58 position per gene per time point) as the expression values in this region vary along the anterior-posterior axis, but is fixed along the dorsal-ventral axis. Choosing the trunk region for our analysis has one further advantage that the key transcription factors acting in this region are well known.

This time series data presents the expression levels of the above mentioned seven gene products at different positions along the anterior-posterior axis. The data is thus a 3D matrix V where the element $v_a^i(t)$ is the expression level of gene a, position i and time t where $a \in \{1, 2, \ldots, 7\}, i \in \{1, 2, \ldots, 58\}$ and $t \in \{1, 2, \ldots, 10\}$. The mapping of our time points with the actual age of the embryo is provided in Table 2-1.

Gene Sequence Data & PWM Data

Scroeder et al. [55] reports 52 cis-regulatory-modules(CRM) within the genomic region of 29 genes of the *Drosophila melanogaster* genome. We selected 10 of them all of which reside in the regulatory region of the four gap target genes *hb, Kr, gt* and *kni.* We pick only those modules which drive expression at the trunk region, i.e.

Table 2–1: Mapping of the time points used in our study with the actual age of *Drosophila*

Time Point	Age of embryo
1	0
2	11
3	24
4	30
5	37
6	43
7	49
8	55
9	62
10	68

35% to 92% region of the A-P axis. The CRMs used in our study are listed in Table 2-2. Figure 2-1, taken from Perkins et al. [42] provides some images of the *Drosophila* embryo at different stages of devlopment. It also shows (Figure 2-1 (I-L)) the gap gene expression at the final time point as a function of A-P position.

The PWMs for the TFBSs of the seven genes mentioned above is taken from the study of Sinha et al. [58] . We have also used the same pseudo-count values as used in the same paper (0.2 for *gt* TFBS, 0.5 for others).

2.1.2 TFBS Strength Values

Schroeder et al. [55] provides the binding site composition of several novel and previously known *cis*-regulatory-modules (CRM) which were detected using an algorithm called Ahab [45] . This CRM-wise binding site composition is presented as a matrix M of integrated profile values, i.e. the fractional occupancy of a site for a given factor, which can be interpreted as the relative binding site *'strengths'* for different TFs. If there are n transcription factors and k different CRMs. $M(i, j)$ is the binding

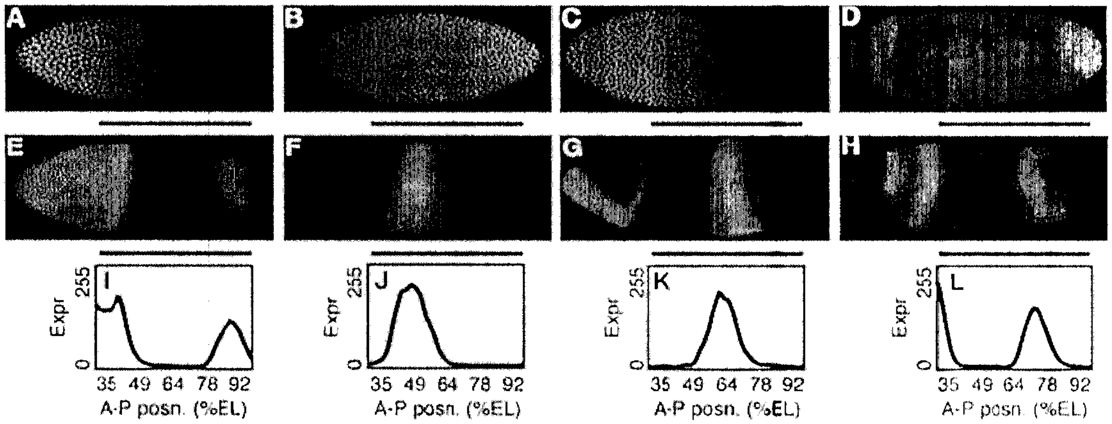


Figure 2–1: Maternal and gap gene expression [42].
(**A** -**C**) *Drosophila* embryos at cleavage cycle 13 fluorescently stained for Bcd (A), Cad (B) and Hb(C) protein. Anterior is to the left and dosral is up.
(**D** -**H**) *Drosophila* embryos at late cleavage cycle 14A fluorescently stained for Tll (D), Hb(E), Kr(F), Kni (G) and Gt(H).
(**I** -**L**) Mean relative gap gene expression as a function of A-P position (measured in percent embryo length) for Hb (I), Kr(J), Kni(K) and Gt(L).

site strength of TF j in module i, where $i \in \{1, 2, \ldots, k\}$ and $j \in \{1, 2, \ldots, n\}$. We compute a cumulative binding site strength (CBSS) matrix B (described below) that presents the total strength of binding sites for each TF-target pair from the given binding site composition matrix M.

Each of these k modules resides in the regulatory region of its respective target. If there are m target genes, the total binding site strength for the TF a for a given target b can be computed by taking the summation of the binding site strengths of a in the CRMs that reside in the regulatory regionof target b. Mathematically,

Table 2–2: The list of CRMs used in our study. **Δgene** denotes the distance to the gene's transcription start site where negative values indicate upstream location and positive indicate downstream.

Target Gene	**Module**	**Size**	**Δgene**
hb	hb_anterior_actv	721	3191
hb	hb_centr_&_post	1023	-3006
Kr	Kr_CD1	1159	-3174
Kr	Kr_CD2_AD1	1707	-1193
gt	gt_(-10)	1745	-8904
gt	gt_(-1)	1239	-163
gt	gt_berman	945	-1815
gt	gt_(-3)	1186	-1410
kni	kni_kd	877	-1177
kni	kni_(+1)	1479	1407

$B(a,b) = \sum_{i=1}^{k} M(i,a),\ P(i,b)$ where,

$$P(i,b) = \begin{cases} 1, & \text{CRM } i \text{ is located in the regulatory (promoter) region of } b \\ 0, & \text{otherwise} \end{cases}$$

We have used both Schroeder et al. [55] data and Sinha et al. [58] for computing the transcription factor binding site strength. Schroeder et al. directly reports the M matrix in their paper, but Sinha et al. does not directly report such a matrix M. However, it provides an efficient algorithm called Stubb [59]. Using this algorithm, we can compute M for all the transcription factors regulating all the targets. We discuss this algorithm in details in the methods section. The CBSS values (the B matrix) derived from Schroeder et al. is given in Table 2-3. The Sinha et al. CBSS values is provided in Table 5-1.

Table 2–3: Cumulative binding site strength matrix of Schroeder et al.

	Hb	**Kr**	**Gt**	**Kni**
Bcd	6.6500	19.9700	18.4000	4.4900
Cad	0	3.9300	13.3000	0
Hb	13.4500	25.3300	14.2400	10.1100
Kr	5.4700	5.5400	21.5000	6.6600
Gt	3.6000	5.5600	0	0
Kni	8.2100	13.2700	12.0300	1.8700
Tll	17.2000	14.4300	44.1200	11.4700

2.2 Methods

In this section, we briefly describe the standard algorithms from the existing literature that we have used in our study. The algorithms that we have devised or customized for our requirements are described in the respective chapters. More specifically, we describe the algorithm devised by Sinha et al. [59, 58] that computes the number of PWM matches (termed as *w-score*) in a given sequence and the standard simulated annealing algorithm which we have used for optimization of different objective functions.

2.2.1 Finding *w-scores*

Sinha et al. [58] defines the notion of *w-score* of a PWM for a given sequence as the number and strength of occurrences of the PWM in the sequence. They have assumed that the sequences of nucleotides in a cis-regulatory module is generated left to right by a stochastic process that successively plants the PWMs of different transcription factors and the background without any overlap. Sinha et al. have assumed that the generation of these sequences follows a Hidden Markov Model(HMM) of order zero that continually picks up and plants a PWM of a TF from the set of

PWMs of all the TFs and the background sequence and thus generates the final nucleotide sequence of the CRM. The selection of HMM of order zero assumes that the choice of a PWM being planted in a given position does not depend on the previously chosen PWMs by the process. Figure 2-2 shows an order zero HMM for a process where there are only 3 states S_0, S_1 and S_2. The transition probabilities associated with these states are p_0, p_1 and p_2 respectively.

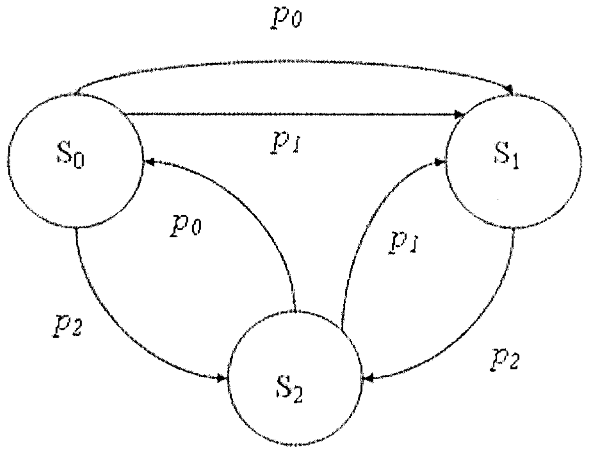


Figure 2–2: An order zero HMM involving 3 states S_0, S_1 and S_2. p_i denotes the transition probability of state i ($1 \leq i \leq 3$).

Given a set of PWMs $W = \{w_0, w_1, w_2, \ldots, w_k\}$, where $w_1, w_2, \ldots, w_k$ represents PWMs for different TFs, and w_0 represents the background frequencies for different nucleotides, the process at any stage may choose any of these $(k+1)$ PWMs to plant. The probability of choosing w_i is p_i. Naturally, $\sum_{i=0}^{k} p_i = 1$.

The Markov process, at any stage, can choose any of the given $(k+1)$ PWMs. The sequence of the PWMs chosen to construct a CRM is called a *parse*. For example, if S_0, S_1 and S_2 has the length 1,2 and 2 respectively, a three nucleotide long CRM can be generated by the parses $\langle S_0, S_1 \rangle, \langle S_0, S_2 \rangle, \langle S_1, S_0 \rangle, \langle S_2, S_0 \rangle$ and $\langle S_0, S_0, S_0 \rangle$. Note that $\langle S_1, S_2 \rangle$ is invalid because it exceeds the sequence length of 3.

For a given sequence S, we can easily compute the posterior probability of any valid parse T of a sequence if we know the HMM parameters $\theta = \{p_0, p_1, p_2, \ldots, p_k\}$. The *w-score* of a PWM w_m given the sequence S and the model paramaeters θ is defined as $\sigma(w_m, S, \theta) = \sum_T X_m(T)P(T|S, \theta)$, where $X_m(T)$ is the number of times w_m occurs in T. Therefore, *w-score* is the expected number of occurence of a TF PWM averaged over all possible configurations(parse) that may be used to generate a given sequence. It can be computed by using the Forward- Backward Algorithm [15] in polynomial time.

2.2.2 Simulated Annealing

Simulated annealing (SA) [34] is a technique that was originally devised for combinatorial optimization problems. Nevertheless, SA has been widely used for all types of optimization problems, for optimization of both discrete and continuous objective functions. It was motivated by an analogy to the statistical mechanics of annealing in solids. Although simulated annealing is not a panacea for every optimization problems on earth, it has been established as a successful method for solving many important classes of optimization problems. It is particularly useful when the problem has many complex constraints, a complex dependency structure among the parameters involved and/or a complex cost function is to be optimized.

To understand the underlying principle of SA, it is useful to look at the process of *annealing* in applied metallurgy for coercing a solid into a low energy crystalline state. The basic idea is to heat the material to a higher temperature to permit the atoms to move freely about. Then, it is carefully and gradually cooled to a lower temperature, so that the atoms can rearrange themselves in such a manner that the material freezes into a perfect and stable crystal. The cooling should not be performed too quickly, otherwise the resulting crystal may not be perfect and stable.

In simulated annealing algorithms in Computer Science, an analogous controlled cooling scheme is utilized for optimization problems. Any optimization problem involves either maximization of an objective function J or minimization of a cost or error function E by changing the solution parameters . In SA, the idea is to make small a random perturbation of the parameters and observe the effect. If this perturbation helps in attaining the goal, i.e. increases J or decreases E, the changes are accepted. If not, the change can still be kept with a probability p such that $0 < p < 1$. With a probability $(1 - p)$, the change is discarded. The value of p is a function of the temperature parameter T. Usually, $p = \exp(\frac{-\Delta E}{k \cdot T})$ where ΔE is the change in error function, i.e. $\Delta E = E_{new} - E_{old}$. k is the Boltzmann constant, which was usually set to a value of 1 in our experiments. It is apparent that for a fixed value of ΔE, p decreases exponentially as we decrease T. Therefore at a higher temperature, a bad move (change) is allowed with a greater probability than that at a lower temperature. The intuition here is that at initial stage, an apparently bad move may be able to lead towards a good final solution, but at later stages of optimization, it is less likely to do that. The temperature parameter T is gradually decreased and the process continues until it converges to a solution which is good enough for the specific purpose.

Ideally, we should expect SA to yield a globally optimal solution, but for most practical problems, we may have to satisfy ourselves with a reasonable local optimum. The advantage of Simulated Annealing over greedy local search and gradient descent based algorithm is that it is less likely to stuck in a local optima, due to its allowance of bad intermediate moves and the stochastic nature of the algorithm. The disadvantage of SA is its long running time to solve the problem. However, the basic SA algorithm can be customized and hybridized to fit into a particular problem and various heuristic methods exists to make it converge faster. Some of the examples of such customization and hybridization approaches can be found at [61, 65, 27, 63, 44]

.

CHAPTER 3
Static Models and Fits

The static modelling of the gene expression data assumes that the final time point itself carries sufficient information for the discovery of the mode and strength of interactions in the gene regulatory network. Although for this particular dataset, this assumption is quite unrealistic (for example, it is not possible to model auto-regulation using static models), we still hoped that static modelling of the data would be a good starting point for our research. We also expected to identify the key regulatory interactions using the static models.

3.1 Model Equation

The general model equation of the static models can be represented as below:

$y_a^i = f(\sum_{b:b\neq a} w_{ba} \cdot v_b^i + \delta_a)$ for all values of a,b and i,where y_a^i is the predicted final time point expression level of gene a at position i, v_a^i is the given final time point expression level of gene a at position i, δ_a is the bias term of gene a such that $f(\delta_a)$ represents the expression level of a given that no other TFs are present, w_{ba} is the regulatory effect of TF b on the target gene a. The function f can be any function. In our study we have chosen two different functions as f. The first one is a simple linear function $f(x) = x$ which assumes that the expression level of a given gene is just a linear function of the expression levels of its TFs. Our second assumption is a more biologically credible one where we assume that f is a sigmoid function, i.e. $f(x) = \frac{1}{1+\exp(-x)}$. We report the results for both the cases.

There is another notable feature in our formulation of the model. The decay and diffusion terms are not considered, so the gene expression is attributed only to the protein synthesis process. As we know the expression profile of the genes at the final time point, we can employ simple linear regression techniques to fit the model by minimizing the squared error between the predicted and given expression profile. Thus we find out the regulatory effects governing the expression profile of the gene network.

3.2 Linear Regression

We consider several different modelling and formulation on which we apply simple linear regression.

I **Simple Linear Regression(SLR) Using Expression Data Only:** As the first step, we try to fit the data to a simple linear model using linear regression. The expression level of each of the gap genes *hb, gt, Kr* and *kni* is assumed to be a simple linear function of the expression of the rest of the genes. For example, the expression level of Hb is considered to be a linear function of the expression level of Bcd, Cad, Kr, Kni, Tll and Gt. Mathematically, the expression level of gene a can be modelled as, $y_a^i = \sum_{b:b \neq a} w_{ba} \cdot v_b^i + \delta_a$, where the symbols have the same meaning as defined at the beginning of this chapter. Fitting the data using this model for each of the four gap genes yields seven weights; six corresponding to how the product of a gene, working as a TF, regulates the expression level of other genes and the other weight corresponding to the baseline expression of the first gene. For example, after fitting a simple linear model for Hb, we get 7 weights, $w_{bcd,hb}$, $w_{cad,hb}$, $w_{Kr,hb}$, $w_{gt,hb}$, $w_{kni,hb}$, $w_{tll,hb}$

and δ_{hb}, where δ_{hb} is the baseline expression of Hb, i.e. the expression level of Hb when no TF is present. Each $w_{g,hb}$, where g is an element of the set of TFs considered, represents how a TF g affects the expression level of Hb. If this weight is positive, we interpret that g acts as an activator for the Hb expression. If the weight is negative, g is identified as a repressor for Hb.

We have to use four separate input matrices for the four target gap genes. The gene expression profile is given for a total 58 positions along the A-P axis. Therefore the input matrix for a gene has 58 rows (each row corresponds to a position along the A-P axis) and 6 columns (each column represents one TF of the target gene). The input matrix for Hb is given below:

$$\begin{pmatrix} v_{bcd}^{1} & v_{cad}^{1} & v_{kr}^{1} & v_{gt}^{1} & v_{kni}^{1} & v_{tll}^{1} & 1 \\ v_{bcd}^{2} & v_{cad}^{2} & v_{kr}^{2} & v_{gt}^{2} & v_{kni}^{2} & v_{tll}^{2} & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ v_{bcd}^{58} & v_{cad}^{58} & v_{kr}^{58} & v_{gt}^{58} & v_{kni}^{58} & v_{tll}^{58} & 1 \end{pmatrix}$$

The output vector of Hb will just be a 58 x 1 matrix, representing the expression level of Hb at the 58 different positions.

II **Single Matrix Regression(SMR) Using Expression Data Only:** Our second model assumes that a TF influences all its target by the same strength (magnitude and mode). This assumption is unlikely to be correct, because although, by most models, an activator in the gap gene system usually activates all its target and an repressor usually represses all its target if we do not consider auto-regulation, there is hardly any evidence that the magnitude of this effects should be the same irrespective of the target gene involved. Nevertheless, this simplistic assumption helps us to reduce the number of parameters to describe the model. The model can be mathematically expressed as : $y^i_a = \sum_{b:b\neq a} w_b v^i_b + \delta_a$.

Instead of using four separate input and output matrices, we combine the input matrices together to obtain a single big input matrix. Before combining the input matrices, a column of zeros corresponding to the expression level of the gene itself is inserted into the matrices. Four new columns corresponding the bias terms for Hb, Kr, Gt and Kni were also added to the input matrix. The output matrix is similarly prepared by combining the observation of the four gap genes in the same order as the input matrix. The input matrix is given on the next page.

$$\begin{pmatrix}
v_{bcd}^{1} & v_{cad}^{1} & 0 & v_{kr}^{1} & v_{gt}^{1} & v_{kni}^{1} & v_{tll}^{1} & 1 & 0 & 0 & 0 \\
v_{bcd}^{2} & v_{cad}^{2} & 0 & v_{kr}^{2} & v_{gt}^{2} & v_{kni}^{2} & v_{tll}^{2} & 1 & 0 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
v_{bcd}^{58} & v_{cad}^{58} & 0 & v_{kr}^{58} & v_{gt}^{58} & v_{kni}^{58} & v_{tll}^{58} & 1 & 0 & 0 & 0 \\
v_{bcd}^{1} & v_{cad}^{1} & v_{hb}^{1} & 0 & v_{gt}^{1} & v_{kni}^{1} & v_{tll}^{1} & 0 & 1 & 0 & 0 \\
v_{bcd}^{2} & v_{cad}^{2} & v_{hb}^{2} & 0 & v_{gt}^{2} & v_{kni}^{2} & v_{tll}^{2} & 0 & 1 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
v_{bcd}^{58} & v_{cad}^{58} & v_{hb}^{58} & 0 & v_{gt}^{58} & v_{kni}^{58} & v_{tll}^{58} & 0 & 1 & 0 & 0 \\
v_{bcd}^{1} & v_{cad}^{1} & v_{hb}^{1} & v_{kr}^{1} & 0 & v_{kni}^{1} & v_{tll}^{1} & 0 & 0 & 1 & 0 \\
v_{bcd}^{2} & v_{cad}^{2} & v_{hb}^{2} & v_{kr}^{2} & 0 & v_{kni}^{2} & v_{tll}^{2} & 0 & 0 & 1 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & & & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & & & \\
v_{bcd}^{58} & v_{cad}^{58} & v_{hb}^{58} & v_{kr}^{58} & 0 & v_{kni}^{58} & v_{tll}^{58} & 0 & 0 & 1 & 0 \\
v_{bcd}^{1} & v_{cad}^{1} & v_{hb}^{1} & v_{kr}^{1} & v_{gt}^{1} & 0 & v_{tll}^{1} & 0 & 0 & 0 & 1 \\
v_{bcd}^{2} & v_{cad}^{2} & v_{hb}^{2} & v_{kr}^{2} & v_{gt}^{2} & 0 & v_{tll}^{2} & 0 & 0 & 0 & 1 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & & & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & & & \\
v_{bcd}^{58} & v_{cad}^{58} & v_{hb}^{58} & v_{kr}^{58} & v_{gt}^{58} & 0 & v_{tll}^{58} & 0 & 0 & 0 & 1
\end{pmatrix}$$

After this pre-processing step, the data is fit to a linear model using linear regression. This linear regression yields eleven weights in total. Seven of these weights (w_{bcd} , w_{cad} , w_{kr} , w_{gt} , w_{kni} , w_{tll} and w_{hb}) denote how a particular TF affects the expression of all other genes. The rest four weights δ_{hb} , δ_{Kr} , δ_{gt} and δ_{kni} denote the baseline expression of Hb, Kr, Gt and Kni respectively. As the number of parameter is much less than the simple linear regression (11 instead of 28), we expect that SMR model to result in more error than the SLR models.

III **Single Matrix Regression (SMR) Using Expression Data and Binding Site Information:** The binding site information is taken into consideration in this step. We assume that the effect of the binding sites for a particular TF is independent of the relative position of the CRMs in which they reside. We also assume that the cumulative binding site strength (CBSS), of a TF for a gene can be represented by the sum of the number of binding sites for that TF in all the modules associated to the gene. We have already defined the notion of CBSS in Chapter 2. Intuitively, CBSS should be a key factor driving the expression level of a gene. If TF a and b has the same magnitude of effect (i.e. same weight) on all the genes and the cumulative binding site strength of TF a is f times the cumulative binding site strength of b, we assume here that for the same concentration of TF a and b, the contribution of a should be f times to the contribution of b. The model can be mathematically expressed as , $y_a^i = \sum_{b:b\neq a} w_{ba} \cdot B(b,a) \cdot v_b^i + \delta_a$. Here $B(b,a)$is the CBSS of TF b in the regulatory region of target gene a.

At first, the CBSS values for all the genes are calculated from the binding site data. The matrix has already been reported in Table 2-3. After calculating the CBSS, i.e. $B(b, a)$ values, we construct the input and output matrix. The output matrix is the same as the output matrix for single matrix regression using expression data only. The input matrix is also constructed in the same manner as single matrix regression. The only difference is that instead of using the expression data directly, we multiply the expression level values of the original matrices with the corresponding cumulative binding site strengths.

We expect that using binding site information should yield better result than the single matrix regression without binding site information if our underlying assumptions are correct. However, if otherwise, the performance of prediction using binding site information should not be good enough.

IV **Regression Without Baseline:** If we assume that the regulation within the network is entirely transcriptional, then a gene will not be expressed in absence of all the TFs. However, in all the three experiments above, the presence of one or more bias term violates this assumption. To test this assumption, we repeat the entire set of experiment without having provision for any baseline/ bias terms.

3.2.1 Performance Evaluation

For linear regression, usually cross validation is used to evaluate the quality of the fit. However, in this case, the data are not remotely i.i.d as each rows corresponds to a spatial position along the A-P axis of the embryo body and so strong correlation is evident between spatially adjacent samples. For this reason we could not use cross validation or test set error for evaluating the fit. We had to rely on training set error, visualizing the fit and prior biological knowledge to evaluate different fits. The training set error is presented in terms of root of the sum of squared error between observed and predicted expression values. For gene a, the root of the sum of squared error is $\sqrt{\sum_i (y_a^i - v_a^i)^2}$. Table 3-5 lists the prediction errors for all the static fit experiments.

3.2.2 Weights Obtained From Linear Regression

The weights for different genes obtained by running different experiments have been listed in Table 3-1 and Table 3-2. There are clearly some trends visible in the weights. The signs of the weights are quite consistent except for Cad which has been termed as a repressor for *Kr* and *kni* using SLR and as an activator for the rest of the experiments. According to current literature [42, 50, 53, 29] Cad, however, should be acting as an activator for all gap genes. Bcd is always termed as a strong activator and Hb, Kr, Gt and Kni as repressors. However, while comparing the relative strength of TFs, it should be remembered that the range of values for Bcd expression is from 0 to about 45 (in the trunk region) while the range of values for other gene expression is from 0 to 255. So Bcd weights should be scaled accordingly for comparison. Even after scaling (not shown), the Bcd weights remain pretty high.

Another point to note is that the weights obtained after binding site consideration are much lower than the other weights. This is expected because as we are multiplying CBSS values to the expression value, the weights need to be much lower to balance for the factor. An interesting result is that when we do not consider the binding site strength, Hb seems to be the strongest repressor. However, when we do consider the binding site strength, Gt emerges out as the strongest repressor. We may interpret that if Gt and Hb had the same number of binding sites for all the genes, then Gt would have the strongest repressing effect. However, as the number of binding site for Hb is greater than that of Gt, Hb has more total contribution to the repression of the genes.

The baseline values for different genes vary greatly for different experiments. Hb shows strong negative baseline for simple linear regression while strong positive baseline for the rest. For Gt, the single matrix regression yields a negative baseline while the others result in positive baseline. For Kni and Kr, the baselines reported are always positive although their magnitude vary a lot.

Table 3–1: Weight matrix obtained from regressions with baseline (The 'X indicates not applicable for the experiment.)

	Hb SLR	**Kr SLR**	**Gt SLR**	**Kni SLR**	**SMR**	**SMR with BS**
Bcd	8.1710	4.3403	7.3935	3.9102	4.7466	0.3651
Cad	1.8972	-0.2568	1.2186	-0.1182	0.1432	0.0907
Hb	X	-1.0898	-1.0639	-1.1405	-1.0747	-0.0607
Kr	-0.4390	X	-0.7361	-0.7739	-0.7695	-0.0364
Gt	-0.5599	-0.9618	X	-0.8406	-0.8663	-0.1516
Kni	-0.4294	-0.7233	-0.6013	X	-0.675	-0.0573
Tll	-0.2618	-0.7115	-0.6311	-0.7571	-0.6838	-0.0196
Hb Baseline	-115.0520	X	X	X	85.3533	74.2320
Kr Baseline	X	138.0452	X	X	100.4638	74.2391
Gt Baseline	X	X	-15.2131	X	94.7049	-7.4818
Kni Baseline	X	X	X	133.1447	101.3379	62.3629

3.2.3 Contributions Plot

In each of the experiments, the prediction of the gene expression level is compared with the actual gene expression level. Moreover, the contribution of each TF is also shown to infer and validate biological results regarding transcription factors.

Table 3–2: Weight matrix obtained from regressions without baseline(The 'X indicates not applicable for the experiment.)

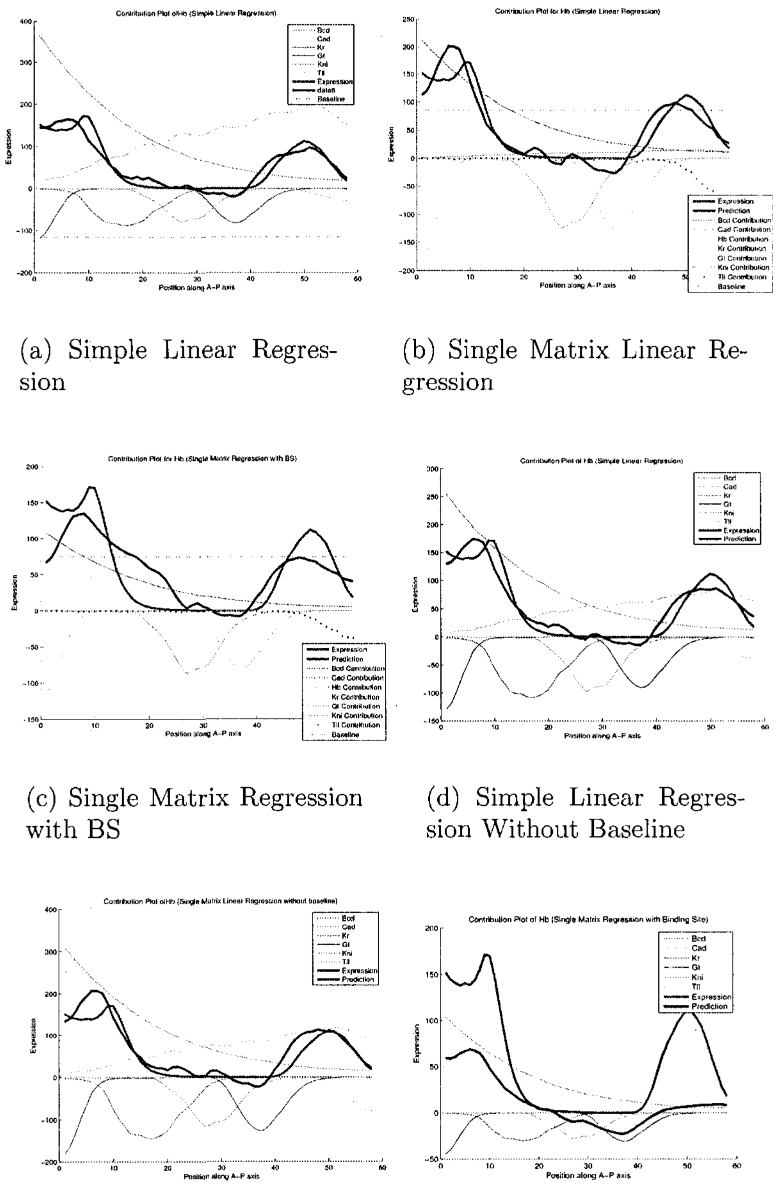
	Hb SLR	**Kr SLR**	**Gt SLR**	**Kni SLR**	**SMR**	**SMR with BS**
Bcd	5.7164	8.4521	7.0003	7.7971	6.9325	0.3501
Cad	0.7967	1.2291	1.0636	1.3124	1.1164	0.0375
Hb	X	-1.2825	-1.0516	-1.3214	-1.0586	-0.0391
Kr	-0.5406	X	-0.7431	-0.7562	-0.7260	-0.0276
Gt	-0.6139	-1.0292	X	-0.8933	-0.8613	-0.0587
Kni	-0.5246	-0.7124	-0.6076	X	-0.6268	-0.0186
Tll	-0.3381	-0.7004	-0.6373	-0.7448	-0.7012	0.0021

Such plots were made for each of the mentioned four gap genes. The results of these experiments for each target gene have been explained below.

I **Hunchback (Hb**) The prediction of the hunchback gene expression according to simple linear regression is shown in the Fig 3-1(a). The prediction is found to be quite close to the actual Hb gene expression observed. The two maternal genes, Bicoid and Caudal act as activators, whereas all the gap genes i.e. Kr, Gt, Kni and Tll act as repressors. The baseline is negative which indicates that Hb will not be expressed if the activators are not present.

The single matrix linear regression gave a very good prediction as well (Fig. 3-1(b)). However, some very interesting differences were noted. The contribution of caudal as an activator is almost negligible. The repressors are much stronger than seen in the case of linear regression for Hb. Hence, the baseline is positive to balance the effect of the strong repression.

The Fig. 3-1(c) shows the result of the single matrix regression when binding strength is used. The peaks seen in the predicted expression do not match with the actual expression level too well and it resulted in much greater training set



(a) Simple Linear Regression

(b) Single Matrix Linear Regression

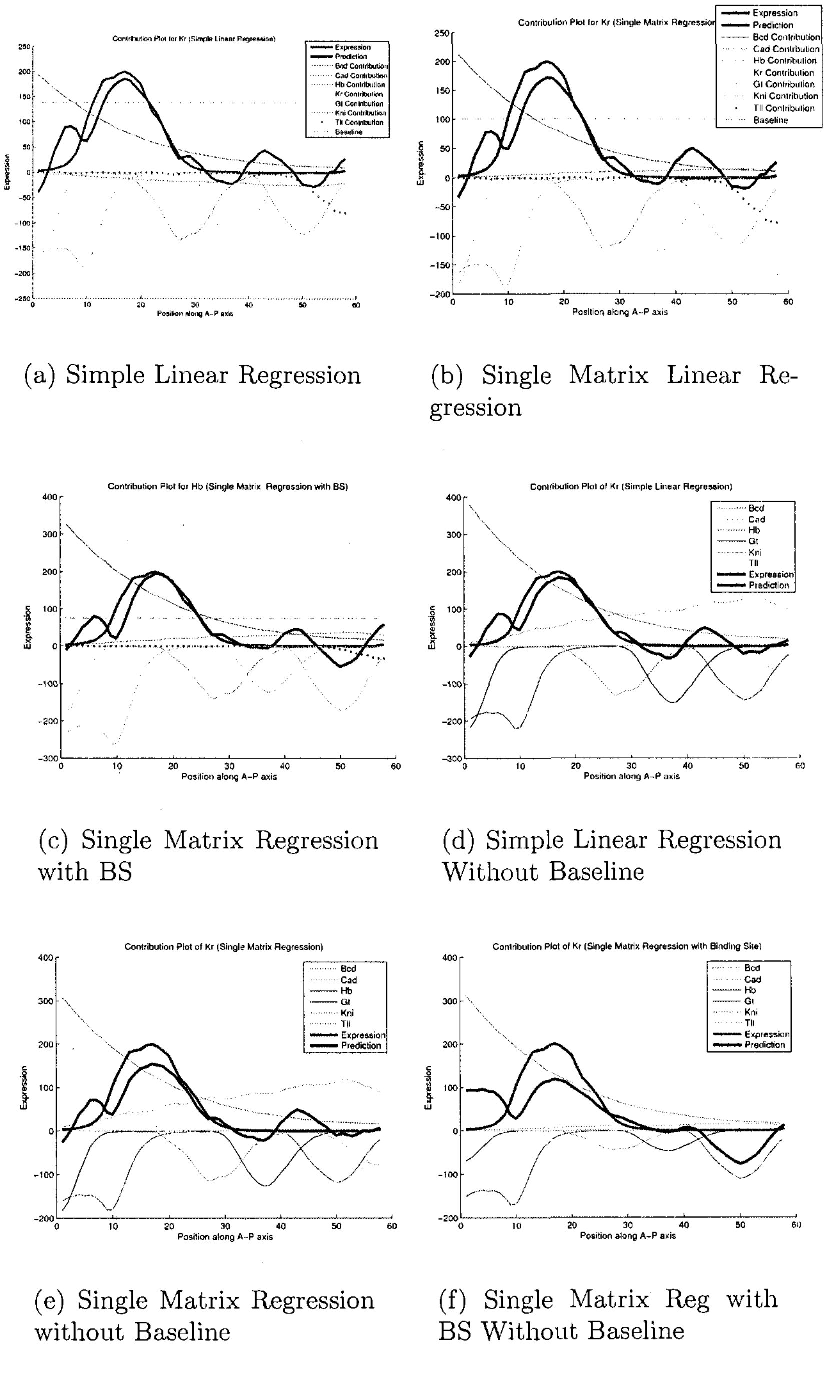
(c) Single Matrix Regression with BS

(d) Simple Linear Regression Without Baseline

(e) Single Matrix Regression without Baseline

(f) Single Matrix Reg with BS Without Baseline

Figure 3–1: Hb Linear regression contributions plots. The thick blue line is the observed expression profile. The thick black line represents the model prediction. Thin lines denote the contributions for each of the TFs.



(a) Simple Linear Regression

(b) Single Matrix Linear Regression

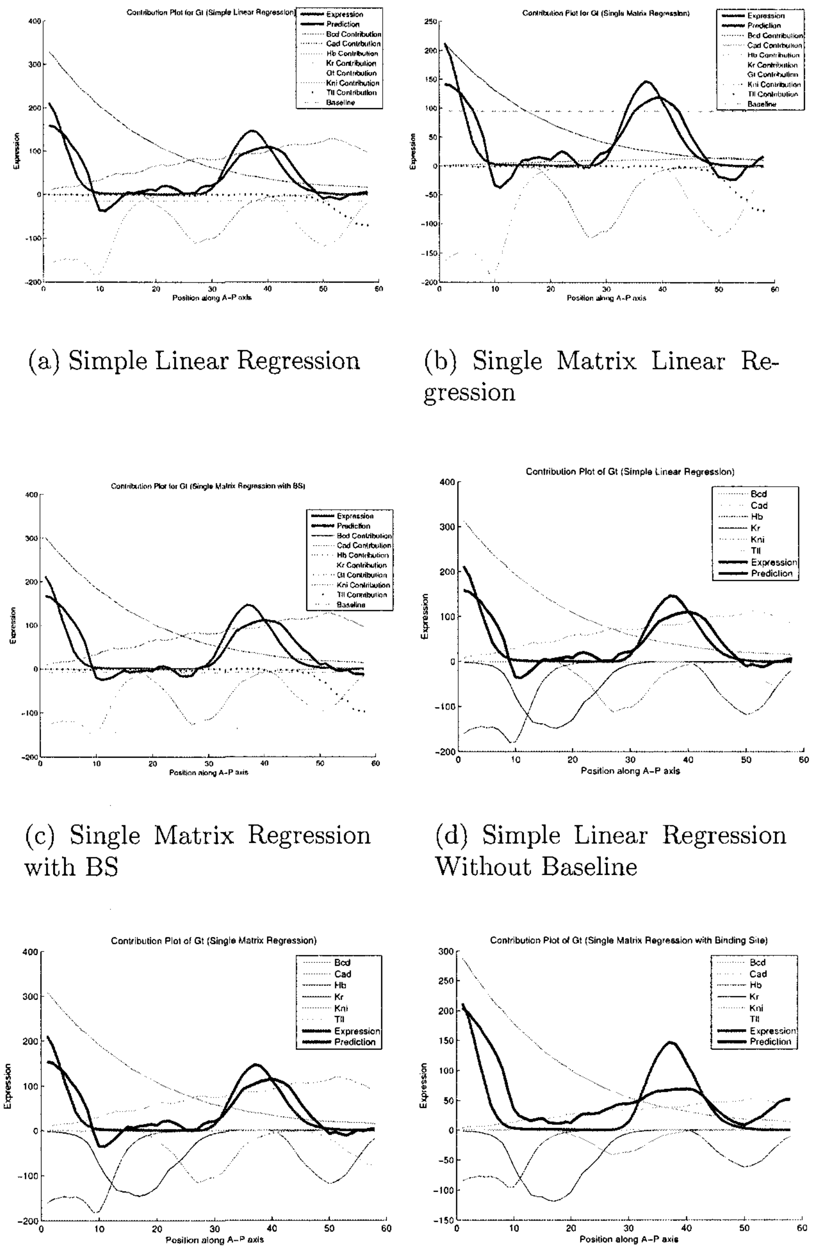
(c) Single Matrix Regression with BS

(d) Simple Linear Regression Without Baseline

(e) Single Matrix Regression without Baseline

(f) Single Matrix Reg with BS Without Baseline

Figure 3–2: Kr Linear regression contributions plots. The thick blue line is the observed expression profile. The thick black line represents the model prediction. Thin lines denote the contributions for each of the TFs.



(a) Simple Linear Regression

(b) Single Matrix Linear Regression

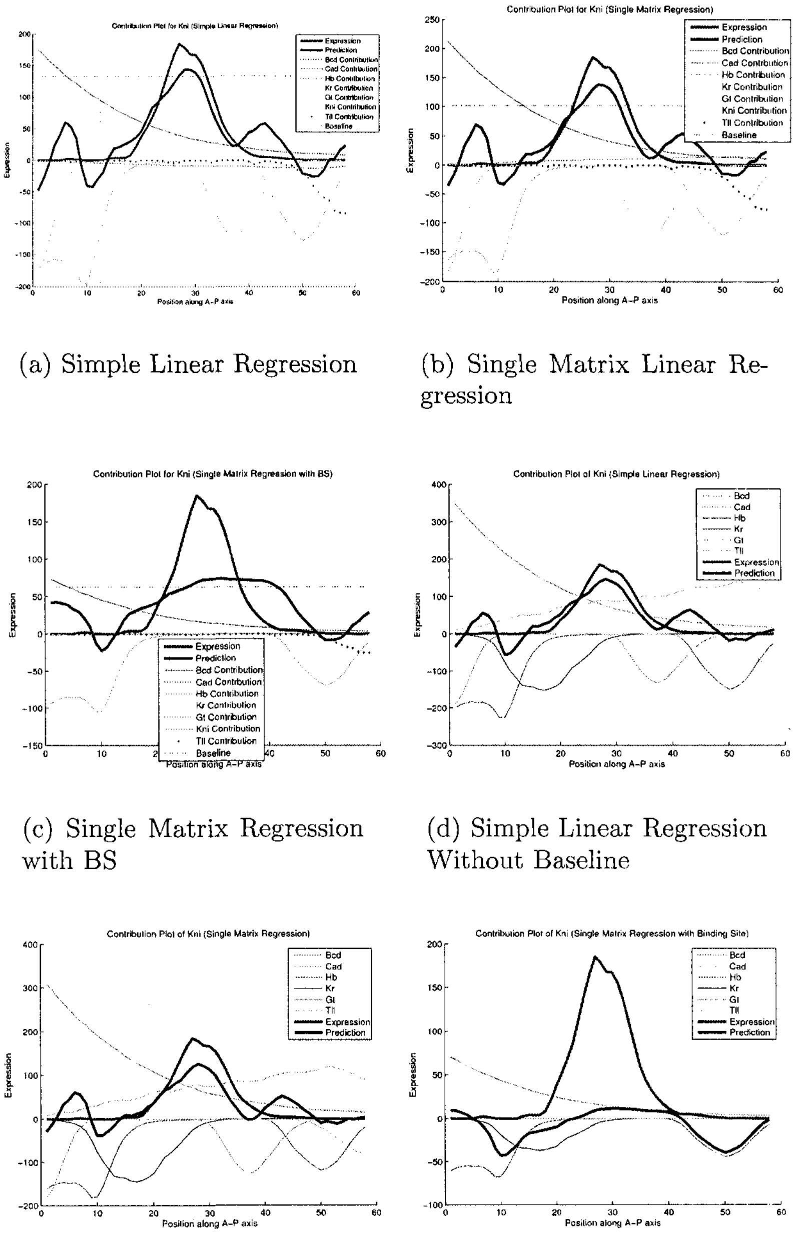
(c) Single Matrix Regression with BS

(d) Simple Linear Regression Without Baseline

(e) Single Matrix Regression without Baseline

(f) Single Matrix Reg with BS Without Baseline

Figure 3–3: Gt Linear regression contributions plots. The thick blue line is the observed expression profile. The thick black line represents the model prediction. Thin lines denote the contributions for each of the TFs.



(a) Simple Linear Regression

(b) Single Matrix Linear Regression

(c) Single Matrix Regression with BS

(d) Simple Linear Regression Without Baseline

(e) Single Matrix Regression without Baseline

(f) Single Matrix Reg with BS Without Baseline

Figure 3-4: Kni Linear regression contributions plots. The thick blue line is the observed expression profile. The thick black line represents the model prediction. Thin lines denote the contributions for each of the TFs.

error. The curve also varies from the actual expression in the middle part of the trunk. These results show that there might be some discrepancies in the binding site strength data.

Both Fig 3-1(d) and 3-1(e) show that the result of simple linear regression and single matrix linear regression are quite encouraging even without the baseline expression. An important point to note in these graphs is that Caudal (Cad) acts as a strong activator. This positive regulation of Cad is supported by Jaeger et al. [29], but not supported by Perkins et al. [42].

According to the Fig 3-1(f), the single matrix linear regression fails to explain the gene expression for Hb when both expression data and binding site strength are used and we do not take the baseline into consideration.

II **Kruppel (Kr):** The first experiment for Kruppel using simple linear regression provided us with a good prediction as shown in the Fig. 3-2(a). The fit is not as good as seen for Hb. There are a few false peaks in the prediction. The interesting thing to note is that Caudal (Cad) acts as a weak repressor here. Single matrix regression gives us similar results (Fig. 3-2(b). The only point of importance is yet again the behaviour of Caudal (Cad) that now acts as a weak activator. Adding the binding site strength for Kruppel (Kr) (Fig. 3-2(c)) does not have a big impact on the prediction. In fact, the prediction is still quite good. Similar results are seen in the case of simple linear regression without baseline (Fig. 3-2(d)) and single matrix linear regression without baseline (Fig. 3-2(e)). However, in all of these predictions the presence of two extra peaks make the prediction bad. The only factor that seems to change is the effect of

Caudal (Cad) gene that acts as a strong activator in the absence of a baseline. The last experiment with both expression data and binding site information (without baseline) did not demonstrate satisfactory prediction performance.

III **Giant (Gt):** Fig. 3-3(a) confirms our idea that the simple linear regression should provide us with the best results in terms of training set error. The prediction is quite close to the gene expression. Both Bicoid and Caudal (Cad) act as strong activator while the gap genes are strong repressors. The single matrix regression also gives a close fit as in Fig. 3-3(b). The low activation strength of Caudal is the only difference seen that is compensated by a much higher baseline.

The experiment for single matrix linear regression with the use of binding site information probably gives us the best result in case of Giant (Fig. 3-3(c)) and Kruppel (Kr) (Fig. 3-2(c)). The behaviour of Caudal (Cad) once again changes as it can now be seen as a strong activator.

The simple linear regression without baseline (Fig. 3-3(d)) and the single matrix linear regression without baseline (Fig. 3-3(e) gave us almost the same results . Caudal (Cad) is a strong activator, which matches with the result of the single linear regression experiment (Fig. 3-3(a)). However, the single matrix regression with baseline (Fig. 3-3(b)) and single matrix linear regression without baseline provide us with contradictory results in terms of the effect of Caudal (Cad) on the gene regulation. In fact, the Caudal (Cad) seems to act as a strong activator to balance the effect of the absence of the baseline.

Figure 3-3(f) solidifies our idea that the single matrix linear regression using binding site information fail to explain the gene regulation if the baseline is not present.

IV **Knirps (Kni):** Simple linear regression gives the best results (Fig. 3-4(a)) also for Knirps. However, the prediction is not as good as seen in the case of other genes. There are two false peaks that are pretty evident. A similar prediction is seen when the expression data is used for the single matrix linear regression for *kni* (Fig. 3-4(b)).

The single matrix linear regression when both the expression data and binding site information is used gives us quite unsatisfactory results. However, these drastic results can be explained by the fact that the binding site strength data for *kni* is not reliable as reported by the Schroeder et. al. [55].

The simple linear regression without the baseline (Fig. 3-4(d)) again gives us similar results as seen in simple linear regression with baseline. However, Caudal (Cad) acts as a strong activator instead of a weak repressor to compensate for the high baseline. Moreover, in this case, Bicoid (Bcd) also is a much stronger activator.

Using single matrix gives similar results with or without the baseline (Fig. 3-4(e)). Once again, the difference is the behaviour of Caudal (Cad), which acts as a strong activator without the baseline. An important thing to note for Knirps is the change in the curve to show a better prediction of the gene expression in the extreme posterior trunk without the baseline. The last experiment (Fig. 3-4(f)) further verifies the poor performance of single matrix regression with binding site strength but without a baseline. The prediction is not significant at all. The results are the worse, perhaps due to the the effect of some missing TFs.

3.2.4 Analysis of Linear Regression Results

The results of the experiments clearly show that the simple linear regression (without binding site information) gives us minimum error predictions in most cases. However, the weights obtained from simple linear regression are sometimes non-conforming to the current belief about the effect of the respective TFs. There is still a lot of room for improvement in the prediction. An interesting observation is that the single matrix linear regression also gives us almost the same prediction results. However, this is done using much fewer parameters.

Another point to notice is that for simple linear regression and single matrix linear regression without baseline, the predictions of the gene expression for the gap genes are once again almost the same. The important feature is the behaviour of the maternal factor Cad which becomes a strong activator in the absence of the baseline. Seemingly, it just balances the effect of the absence of the baseline.

The gene expression get much worse when the binding site strength is used for *hb* and *kni*. Thus, adding binding site strength does not improve the result on the single matrix linear regression. Hence, it seems quite probable that the binding site strength data is not reliable in all cases. Some of the calculated strengths are not accurate and some of the modules may be missing. Moreover, we have just used the sum of the binding site strengths for different modules for each gene, which may not be true. The binding site strength of each TF might be a more complex function of the binding site strength of each module.

Finally, using binding site data without baseline gives us very poor results in all cases. It shows that the linear models cannot incorporate binding site information without the aid of a bias term which vertically shifts the prediction just to adjust the error term without changing the shape of the final output.

In some of the previous related studies [55, 50] , Hb has been termed as an activator in some specific conditions. According to our experiments, we did not notice any such case. Hunchback was seen to be a strong repressor in most cases.

3.3 Logistic Regression

We consider several different modelling and formulation of logistic regression, similar to what we have considered for linear regression.

I **Simple Logistic Regression (SLR) Using Expression Data Only:**

Mathematically, the expression level of gene a can be modelled as,

$y_a^i = g(\sum_{b:b\neq a} w_{ba} v_b^i + \delta_a)$

where g is a logistic function $g(x) = \frac{1}{1+exp(-x)}$. The other symbols have the same meaning and interpretation as defined previously in this chapter. A notable difference is that the output matrix is normalized between 0 and 1 by dividing each of its entries by the highest expression level seen at the column.

We expect the simple logistic regression to work better than the simple linear regression case as the logistic function presents a more biologically realistic solution for modelling smooth threshold-based activation.

II **Single Matrix Logistic Regression (SMR) Using Expression Data Only:** This model assumes that a TF influences all its target by the same strength (magnitude and mode). The model can be mathematically expressed as : $y_a^i = g(\sum_{b:b\neq a} w_b v_b^i + \delta_a)$

The input and output matrices are constructed the same way as in the single matrix linear regression case, the only difference is that the output matrices are normalized.

Single matrix regression is expected to yield in more error than the simple logistic regression, due to the use of less parameters to define the model.

III **Single Matrix Logistic Regression (SMR) Using Expression Data and Binding Site Information:** The model can be mathematically expressed as: $y_a^i = g(\sum_{b:b\neq a} w_{ba} \cdot B(b,a) \cdot v_b^i + \delta_a)$.

The input and output matrices are constructed in the same way as in single matrix linear regression with binding sites.

IV **Regression Without Baseline:** We repeat the entire set of logistic regression experiments without having provision for any baseline/ bias terms.

3.3.1 Performance Evaluation

We use the training set error, visualization of the fits and prior biological knowledge as the means to evaluate different fits.

3.3.2 Weights Obtained From Logistic Regression

The weights for different genes obtained by running different experiments have been listed in Table 3-3 and Table 3-4. The weights are obtained using standard matlab function *fminsearch.* The signs of the weights in general match with our

Table 3–3: Weight matrix obtained from logistic regressions with baseline (The 'X indicates not applicable for the experiment.)

	Hb SLR	**Kr SLR**	**Gt SLR**	**Kni SLR**	**SMR**	**SMR with BS**
Bcd	0.1003	0.3370	0.0771	0.6079	0.2248	0.0178
Cad	0.0117	-0.0690	-0.0498	-0.0319	0.0043	0.0047
Hb	X	-0.0396	-0.0304	-0.1886	-0.0546	-0.0032
Kr	-0.0132	X	-0.2803	-0.0512	-0.0326	-0.0029
Gt	-0.0153	-0.4504	X	-0.0183	-0.0395	-0.0063
Kni	-0.4182	-0.0068	-0.0099	X	-0.0267	-0.0024
Tll	-0.0238	0.0123	-0.0583	0.0114	-0.0207	-0.0008
Hb Bias Term	0.1239	X	X	X	-0.0039	-0.8258
Kr Bias Term	X	0.6882	X	X	0.4243	-5.2904
Gt Bias Term	X	X	4.4302	X	-0.3131	-0.0694
Kni Bias Term	X	X	X	-0.0287	0.6588	0.1679

findings in linear regression. Bicoid is always termed as an activator and the gap genes always repress each other. Unlike the linear regression case, Tailless is termed as an activator for some of the experiments. However, the current literature is divided upon this issue [42, 29]. Like the linear regression results, Hb seems to be the strongest repressor if we do not consider binding site strengths. But when we consider the binding site strength, Gt emerges to be the strongest repressor. The baseline values for different genes vary greatly for different experiments.

3.3.3 Contributions/ Absence Plot

In case of logistic regression, the contributions of the transcription factors on a target can be more clearly visualized by analyzing the effect of the absence of one TF at a time on the expression level of the target gene. We have calculated and plotted the result of the absence of a TF on the target and thus we determine how a TF influences the expression of a target gene. For each of the experiments, the

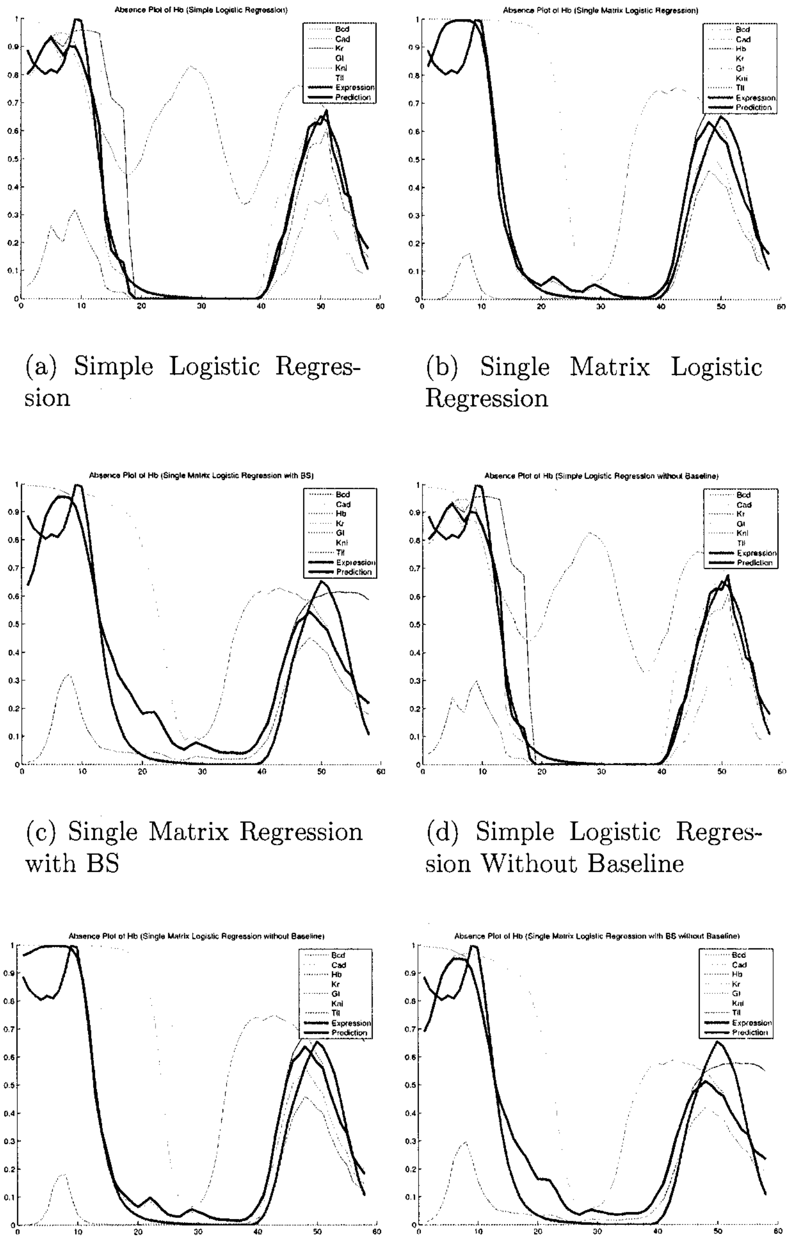
Table 3–4: Weight matrix obtained from logistic regressions without baseline(The 'X indicates not applicable for the experiment.)

	Hb SLR	Kr SLR	Gt SLR	Kni SLR	SMR	SMR with BS
Bcd	0.1033	0.3766	0.2059	0.6070	0.2302	0.0181
Cad	0.0128	-0.0682	-0.0021	-0.0322	0.0035	-0.0002
Hb	X	-0.0418	-0.0425	-0.1885	-0.0566	-0.0057
Kr	-0.0132	X	-0.2083	-0.0512	-0.0311	-0.0028
Gt	-0.0154	-0.4712	X	-0.0183	-0.0328	-0.0059
Kni	-0.4156	-0.0055	-0.0111	X	-0.0266	-0.0026
Tll	-0.0236	0.0171	-0.0826	0.0113	-0.0191	0.0007

predicted and the actual expression level is also shown. The results are presented and analyzed below:

I **Hunchback (Hb):** The prediction of the hunchback gene expression according to simple logistic regression is shown in the Fig 3-5(a). As expected, the prediction matches pretty closely to the actual Hb gene expression observed. Bicoid is the activator for the anterior domain and Caudal is the principal contributor to the formation of the posterior domain. Knirps act as the strongest repressor which inhibits Hb expression in between these two domains. The root squared error for this experiment is less than any of those observed in case of linear regression (Table 3-5) which proves that the threshold based activation of the logistic function provides a better modelling approach in this case .

The single matrix logistic regression gave slightly worse prediction although the error is less than any of the errors observed in linear regression (Fig. 3-5(b)). However, some very interesting differences were noted. All the repressors play an important role in this case. Moreover Caudal is now a less stronger activator



(a) Simple Logistic Regression

(b) Single Matrix Logistic Regression

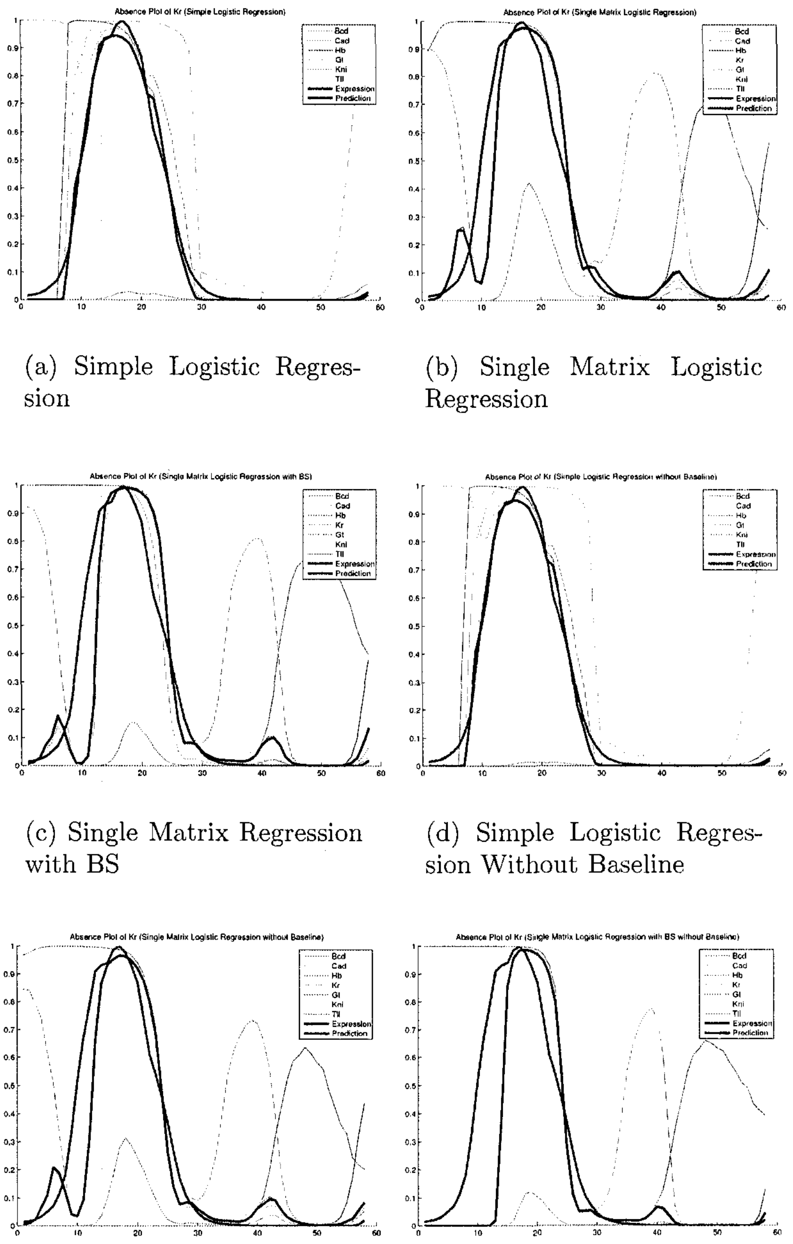
(c) Single Matrix Regression with BS

(d) Simple Logistic Regression Without Baseline

(e) Single Matrix Regression without Baseline

(f) Single Matrix Reg with BS Without Baseline

Figure 3–5: Hb Logistic regression absence plots. The thick lines denote the original and the predicted expression. The thin lines visualize the effect of the absence of each TF.



(a) Simple Logistic Regression

(b) Single Matrix Logistic Regression

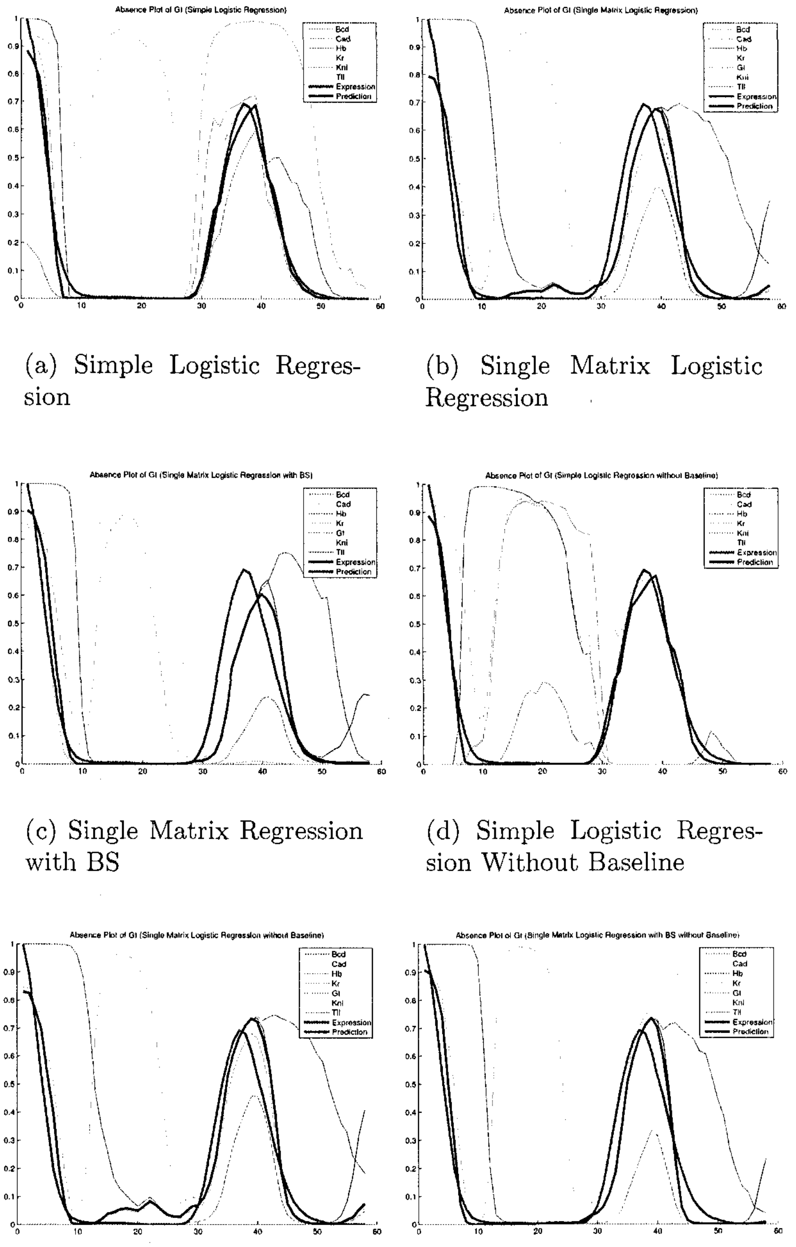
(c) Single Matrix Regression with BS

(d) Simple Logistic Regression Without Baseline

(e) Single Matrix Regression without Baseline

(f) Single Matrix Reg with BS Without Baseline

Figure 3–6: Kr Logistic regression absence plots. The thick lines denote the original and the predicted expression. The thin lines visualize the effect of the absence of each TF.



(a) Simple Logistic Regression

(b) Single Matrix Logistic Regression

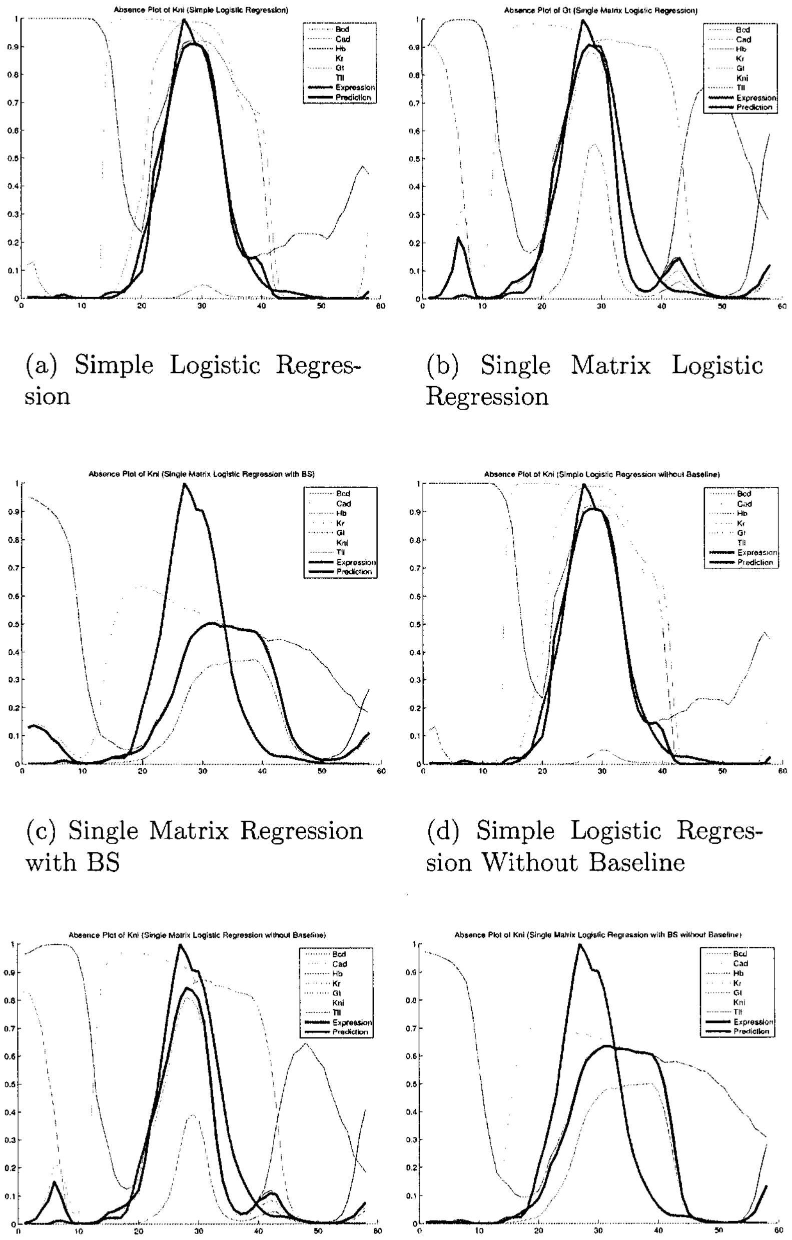
(c) Single Matrix Regression with BS

(d) Simple Logistic Regression Without Baseline

(e) Single Matrix Regression without Baseline

(f) Single Matrix Reg with BS Without Baseline

Figure 3–7: Gt Logistic regression absence plots. The thick lines denote the original and the predicted expression. The thin lines visualize the effect of the absence of each TF.



(a) Simple Logistic Regression

(b) Single Matrix Logistic Regression

(c) Single Matrix Regression with BS

(d) Simple Logistic Regression Without Baseline

(e) Single Matrix Regression without Baseline

(f) Single Matrix Reg with BS Without Baseline

Figure 3–8: Kni Logistic regression absence plots. The thick lines denote the original and the predicted expression. The thin lines visualize the effect of the absence of each TF.

Table 3–5: Prediction error chart for different experiments. The lowest error for each column is marked with bold and the second best with italic.

Experiment	**Hb**	**Kr**	**Gt**	**Kni**
SLR (Linear)	133.77	210.77	184.40	218.02
SMR(Linear)	172.79	219.55	191.94	224.77
SMR(Linear) with BS	253.03	251.72	208.37	334.45
SLR(Linear) (no baseline)	140.95	217.11	184.48	223.69
SMR (Linear)(no baseline)	185.05	238.82	185.66	245.48
SMR (Linear) with BS(no baseline)	436.51	410.30	314.47	510.01
SLR(Logistic)	**73.83**	**59.53**	**53.18**	**29.26**
SMR(Logistic)	114.59	208.49	107.83	81.56
SMR(Logistic) with BS	172.73	256.78	147.37	198.91
SLR(Logistic) (no baseline)	*74.16*	*60.38*	*54.43*	*29.29*
SMR(Logistic)(no baseline)	129.33	225.41	117.55	85.10
SMR(Logistic) with BS(no baseline)	161.88	359.82	104.07	186.63

and the posterior peak formation is attributed to principally Bicoid dependent activation.

Fig. 3-5(c) shows the result of the single matrix regression when binding strength is used. The error does not increase as much as its linear regression counterpart. Still the error is more than the other logistic regression results. Both Fig 3-5(a,d) and 3-5(b,e) show that the result of simple linear regression and single matrix linear regression are very close to what we obtain from regressions with baseline. According to the Fig 3-5(f), even when both expression data and binding site strength are used and we do not take the baseline into consideration, the prediction is still pretty good, which contrasts its linear regression counterpart result.

II **Kruppel (Kr):** The first experiment for Kruppel using simple logistic regression provided us with a very good prediction as shown in the Fig. 3-6(a). Bicoid is the only activator. Interestingly, Caudal has been identified as a strong repressor that represses the expression of Kr in the posterior domain, which does not conform to the literature. Similar results are obtained for simple logistic regression without baseline.

Single matrix regression gives a much greater error which is comparable to the linear regression result. A small false peak is predicted at the posterior end. All the gap genes are acting as strong repressor in this case. Similar results are seen in the case of single matrix linear regression without baseline (Fig. 3-6(e) for which the prediction error is greater than its equivalent linear regression counterpart. When the binding site information is used, the error increases. Especially when no baseline is used, the prediction error increases much and the predicted expression is shifted towards the posterior.

III **Giant (Gt):** For the simple logistic regression (Fig. 3-7(a)), the prediction is pretty well. Bicoid (Bcd) activates the anterior while the posterior is mostly activated by the baseline expression. Kr is the strongest repressor which suppresses the mid-domain expression of gt. The single matrix regression (Fig 3-7(b)) introduces a greater level of error. It marks Cad as an activator. The binding site information does not cause a remarkable increase to the error as in linear regression(Fig 3-7(c)). In fact when there is no baseline, the corresponding error after introducing binding site information is even less than the error observed when binding site information is not used, a phenomena which is unique in this case only.

IV **Knirps (Kni):** For Knirps, the simple logistic regression gives the best of all the results (Fig. 3-8(a)). However, caudal is once again termed as a strong repressor which violates the literature findings. Bicoid is the primary activator and all the gap genes act as repressors for knirps. Similar results are evident when the baseline is not used.

The single matrix logistic regression (Fig 3-8(b)) shows two false peaks at the anterior and posterior. The prediction error is much greater than for the simple logistic case. The results do not change much when we remove the baselines (Fig 3-8(d,e)). But once we add the binding site information, the error increases greatly both with or without baseline (Fig 3-8(f))perhaps due to the already unreliable binding site data for Knirps [55] coupled with the effect of some missing TFs.

3.3.4 Analysis of Logistic Regression Results

The results of the experiments clearly show that the simple logistic regression (without binding site information) with baseline gives us minimum error predictions in all the cases. However, the weights (and contributions) obtained from these regression are sometimes non-conforming to the current belief about the effect of the respective TFs. Single matrix regression increases the training error, although the results are often better than most linear regression results. Binding site data does not usually cause notable deterioration of prediction performance with the exception of the case of *kni* as target, for which the binding site information as provided in [55] lacks input from two transcription factors.

The performance of regression without baseline is usually very good, which suggests that the static data model using logistic function can successfully model the biological observations almost entirely by transcriptional regulation activities of the known transcription factors. The usual increase of error when BS information is introduced may be attributed to i) lack of reliability of the current BS data or ii) Failure of our assumption of a simple additive model while calculating the cumulative BS strength (CBSS) values.

3.4 Summary of Static Fit Results

Our deductions from the static fit results are:

i The static models, although not the most biologically sound one, perform pretty well in predicting the gene expression values. However, the annotation of activators and repressors does not always conform with the current literature results.

ii The bias term plays an important role in case of linear regression. If the bias is positive, usually the weight for Cad is predicted as zero or negative but if no bias is allowed, Cad is usually termed as an activator. When bias is negative, often one or more repressors are predicted to have weaker repressing effect.

iii Logistic (sigmoid) function models the data better than a simple linear function.

iv Single matrix logistic regression performance suggests that the assumption of the similar effect of every TF on all the targets is most likely to be incorrect. The single matrix linear regression performance, however, does not invalidate the assumption.

v Using binding site data introduces much greater errors in linear regression. But in logistic regression, the binding site data performs much better. As logistic modelling seems to be the better of the two models, we can guess that the binding site data provided by [55] is not absolutely unrealistic, although there also seems to be some problems, especially for Knirps CRMs.

In the next chapter, we present the results obtained from fitting dynamical models to the data.

CHAPTER 4
Dynamical Models and Fits

Development is a dynamical process, so it is natural to formulate and fit dynamical models, as opposed to the static models of the previous chapter. Dynamical modelling is, in fact, the most common modelling approach towards modelling of the gene regulatory network (GRN) of *Drosophila melanogaster.* The principal underlying assumption is that the expression level of a particular gap gene at a point of time is dependent on the instantaneous expression profile of all the TFs governing its expression at the immediately previous time point. Using dynamical models, it is possible to model auto regulation within the network. As we have discussed in Chapter 1, the usual approach in the literature is to use a variant of differential equation modelling for the data. However, one common problem is the computational overhead of finding a good fit of the data, principally due to the inherent complexity of the model itself. In our work, we employ a simpler discrete-time gene circuit model to fit the data. The discrete-time model fits are more accurate and the identified regulatory interactions generally conform to the literature. From the static fit results, it became apparent that the sigmoid function based modelling provides a better representation of the biological process involved. Therefore our dynamical models consider only sigmoid based logistic functions for production modelling.

4.1 Discrete-Time Gene Circuit Model

The discrete-time gene circuit model represents the expression profile of a gene at each discrete time point as a function of the expression level of all its TF at the immediately previous time point. The expression profile is thus assumed to be a discrete function over time as opposed to the differential equation model which assumes the expression profile as a continuous function over the time domain. There are two different variants of discrete time models, namely

I Transition-based model, and

II Trajectory-based models.

In our project, we have explored both these models.

4.1.1 Transition-Based Modelling

The model equation can be presented as below:

$$y_a^i(t+1) = R_a \cdot g\left(\sum_b w_{ba} \cdot v_b^i(t) + h_a\right) + (1-\lambda_a)v_a^i(t) \tag{4.1}$$

Where,

$y_a^i(t)$= Predicted expression level of gene a at position i and time t.

$v_a^i(t)$= Observed expression level of gene a at position i and time t.

R_a = Maximum production rate of gene a.

w_{ba} =Regulatory weight of TF b on target gene a.

λ_a = Decay rate of gene a.$(0 < \lambda_a < 1)$

h_a= Bias term for target gene a.

$g(u) = \frac{1}{1+e^{-u}}$

Equation (4.1) models each transition of the gene expression profile, but it does not model the whole trajectory of expression, i.e. only the individual state transitions are modelled, but the predicted output is not used for determining the next time step prediction. Figure 4-1 schematically represents the transition-based models. The objective of the model fitting procedure is to minimize the prediction error term, $E = \sqrt{\sum_{\forall a,i,t}(y_a^i(t) - v_a^i(t))^2}$.

In Equation (4.1), there are actually two terms at the right hand side. The first term is the production term and the second term is the left over from the previous time point expression of gene a after considering an exponential decay of the magnitude $\lambda_a \cdot v_a^i(t)$. Note that we did not model diffusion in this case. The w_{ba} parameters represent the inter-dependency between different genes in the GRN. The interpretation of these w_{ba} parameters are the same in Chapter 3, i.e. if w_{ba} is positive, it indicates that b is an activator of a and a negative value of w_{ba} indicates that b represses a. The higher the magnitude of w_{ba}, the greater is the strength of the regulatory effects for either activation or repression. We have fixed h_a to -3.5 for all the targets in accordance with [42, 29] . For the *Drosophila* gap gene data set, Bcd, Cad and tll are purely transcription factors for the gap genes *hb*, *Kr*,*gt* and *kni*. These four gap genes are acting as both TFs and targets. So the w matrix will have a dimension 7 by 4 as there are 7 TFs regulating 4 gap genes.

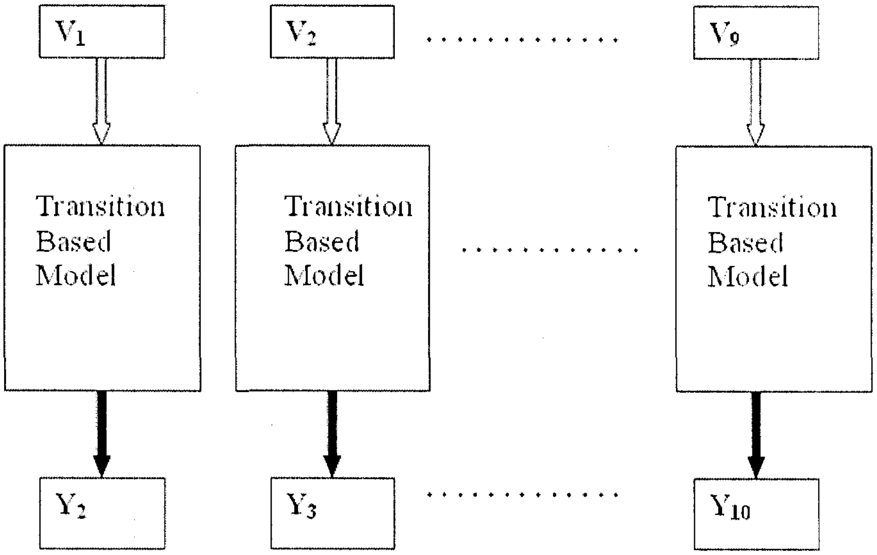


Figure 4–1: The inputs and outputs of the Transition-based models. The filled lines denote outputs and the hollow lines denote inputs to the model. $V_1, V_2, \ldots V_9$ are the observed expression profile matrices for time point 1 through 9. $Y_2, Y_3, \ldots Y_{10}$ are the predicted expression profile matrices for time point 2 through 10.

4.1.2 Trajectory-based Modelling

The model equation can be presented as below:

$$y_a^i(t+1) = D\left(R_a g(\sum_b w_{ba} y_b^i(t) + h_a) + (1-\lambda_a) y_a^i(t)\right) \tag{4.2}$$

where,

$y_a^i(1) = v_a^i(1)$

$y_a^i(t)$= Predicted expression level of gene a at position i and time t.

$v_a^i(t)$= Original expression level of gene a at position i and time t.

R_a = Maximum production rate of gene a.

w_{ba} =Regulatory weight of TF b on target gene a.

λ_a = Decay rate of gene a $(0 < \lambda_a < 1)$.

h_a= Bias term for target gene a.

$g(u) = \frac{1}{1+e^{-u}}$ D = Diffusion Operator

Trajectory-based modelling models the whole trajectory of the gene expression profile, i.e. the predicted expression values at time t are used to predict the expression values at time $(t + 1)$. Figure 4-2 schematically represents the transition-based models. The objective of the model fitting procedure is to minimize the prediction error term, $E = \sqrt{\sum_{\forall a,i,t}(y_a^i(t) - v_a^i(t))^2}$.

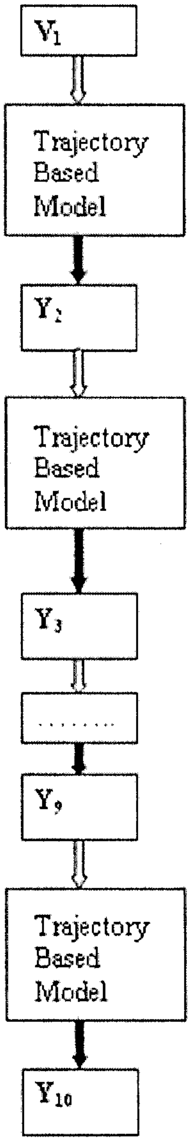


Figure 4–2: The inputs and outputs of the Transition-based models. The filled lines denote outputs and the hollow lines denote inputs to the model. V_1 is the observed expression profile matrices for time point 1 . $Y_2, Y_3, \ldots Y_{10}$ are the predicted expression profile matrices for time point 2 through 10.

For our dataset, the first two time steps are significantly larger (almost double) than the other time steps. For this reason, we have added two fictitious time steps (which we call time step 1.5 and 2.5). Therefore the expression profile at time step 1 is used for simulating the expression profile for time step 1.5 which is in turn used for simulating the expression profile for time step 2. The same procedure is followed for time step 2.5 as well. For the later time steps, predicted expression values at time t are used to predict the expression values at time $(t+1)$. Equation (4.2) models the expression of a gene using the contributions from production, decay and diffusion. The production term is contingent on the expression profiles of all the transcription factors and the w_{ba} parameters representing the inter dependency between different genes in the GRN. The h_a values were fixed to to -3.5 for all the targets in accordance with [42, 29] . The w matrix has a dimension of 7*4 as there are 7 TFs regulating 4 gap genes.

The decay term, just as the transition-based optimization case, is dependent on the previous step expression level of the target gene itself. Diffusion is modelled by convoluting the predicted expression with a discretized gaussian blurring function. The blurring function has 9 components as it has been assumed that only four adjacent nuclei at both side of a nuclei contribute towards the diffusion process at that nuclei. The length of the blurring window and the blurring function vector is chosen empirically such that a rectangular box like expression profile, when convoluted with the particular vector, shows similar spatial slope to the mean slope of the original expression profile of the gap genes. The anterior AP axis border is assumed to be reflecting and te posterior AP axis border is assumed to be absorbing when the convolution is applied. The blurring vector is reported below.

$$\begin{pmatrix} 0.0630 \\ 0.0929 \\ 0.1226 \\ 0.1449 \\ 0.1532 \\ 0.1449 \\ 0.1226 \\ 0.0929 \\ 0.0630 \end{pmatrix}$$

4.2 Transition-Based Modelling Results

4.2.1 Model Fitting

The parameters for this model were obtained by the process of simulated annealing (SA). We have used SA instead of any gradient based optimization method because our empirical studies have shown that the gradient based methods (such as logistic regression) are much prone to getting stuck to a local minima for this particular data set for transition-based optimization.

The magnitudes of the weights (w parameters) were initialized with small random numbers. The sign of the weights were initialized with the signs obtained by the UNC GC fits of Perkins et al. [42]. The decay rate (λ_a) and maximum production rate (R_a) parameters were also initialized to the normalized values obtained by [42] . To ensure that λ_a parameters always remain within the the limit $[0, 1]$, we introduced a function θ such that, $\lambda_a = \frac{1}{1+exp(-\theta_a)}$. On the limit, if θ_a goes to ∞ λ_a goes to 1. If θ_a goes to $-\infty$, λ_a goes to 0. In this way, we can use unconstrained optimization and yet be able to constrain the λ_a values within the allowable range.

The objective of the model fitting procedure is to minimize the prediction error term, $E = \sqrt{\sum_{\forall a,i,t}(y_a^i(t) - v_a^i(t))^2}$. Starting from this initial guesses about the parameters, SA algorithm updates the parameters with specified step size (10^{-3} for w, 5 x 10^{-3} for R, 10^{-4} for θ parameters), and check whether the update process could reduce the error E. If yes, it kept the changes. Otherwise, the probability of keeping the change depended on the temperature of the optimization procedure. The detailed description of SA procedure has already been given in Chapter 2.

4.2.2 Obtained Parameter Values

Table 4-1 lists the weights that minimizes the error function. The other parameters are listed in table 4-2. There are some significant differences in the weights from the weights obtained by the static fits. The key differences are:

i Due to the inherent properties of the static fits, it was not possible to model autoregulation of the gap genes. However, in the case of dynamical models, autoregulation is modelled. Our model finds autoactivation for all the gap genes although for Knirps, the magnitude of the auto-regulatory weight is small.

ii Bicoid and Caudal are always termed as activators. In case of static fits, Caudal was sometimes identified as a repressor.

iii Gap genes are sometimes identified as activators for other gap genes which was never in case of static fit results. However, in most cases such activating weights are small in magnitude.

Table 4–1: Weight matrix obtained from the transition-based fits

	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.0782	0.2741	0.0442	0.5516
Cad	0.0043	0.0122	0.0133	0.0030
Hb	0.0330	-0.1473	0.0024	-0.5222
Kr	-0.0086	0.2637	-0.0447	-0.0213
Gt	0.0149	-0.0272	0.0223	-0.0249
Kni	-0.0678	-0.1948	0.0062	0.0054
Tll	0.0051	-0.8137	-0.0182	-0.1398
Baseline	-3.5	-3.5	-3.5	-3.5

4.2.3 Contribution/Absence Plots

The absence plots are generated the same way they were generated in static fit experiments. Figure 4-3 to 4-7 present these plots. The blue thick line represents the original expression and the black thick line represents the expression levels as predicted by the models. The impact of the absence of a TF is visualized by the thin coloured lines. We analyze the plots for the four target genes.

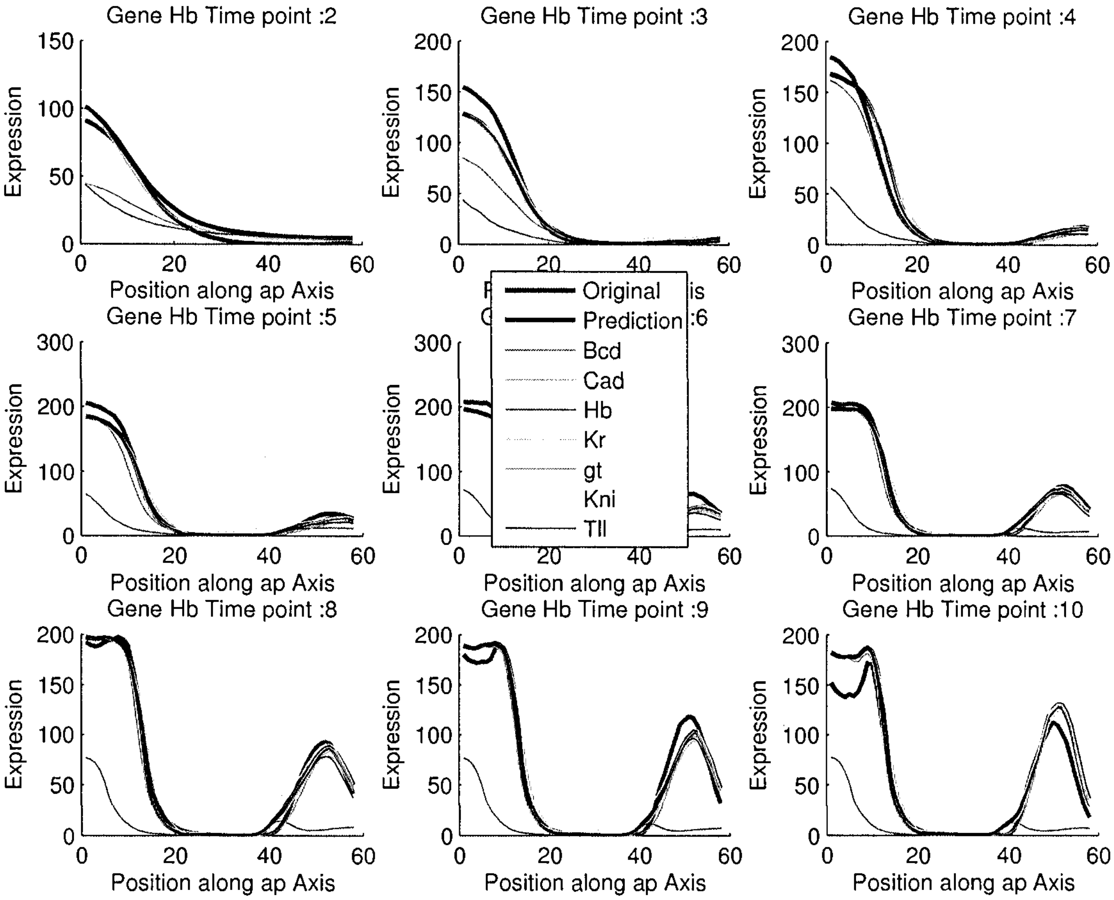


Figure 4–3: Transition-based optimization contribution plots for Hb. The thick lines denote the original and the final expression profiles. The thin lines represent the effect of the absence of each TFs.

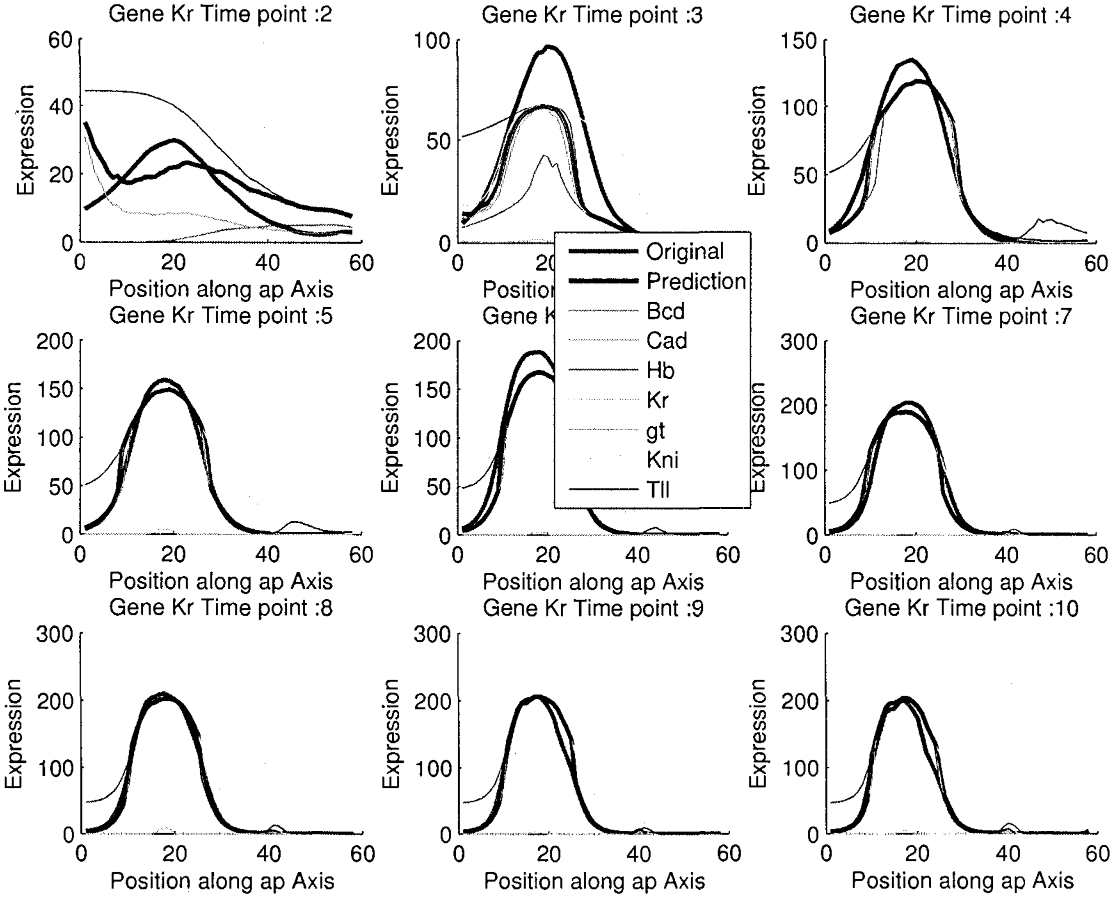


Figure 4–4: Transition-based optimization contribution plots for Kr. The thick lines denote the original and the final expression profiles. The thin lines represent the effect of the absence of each TFs.

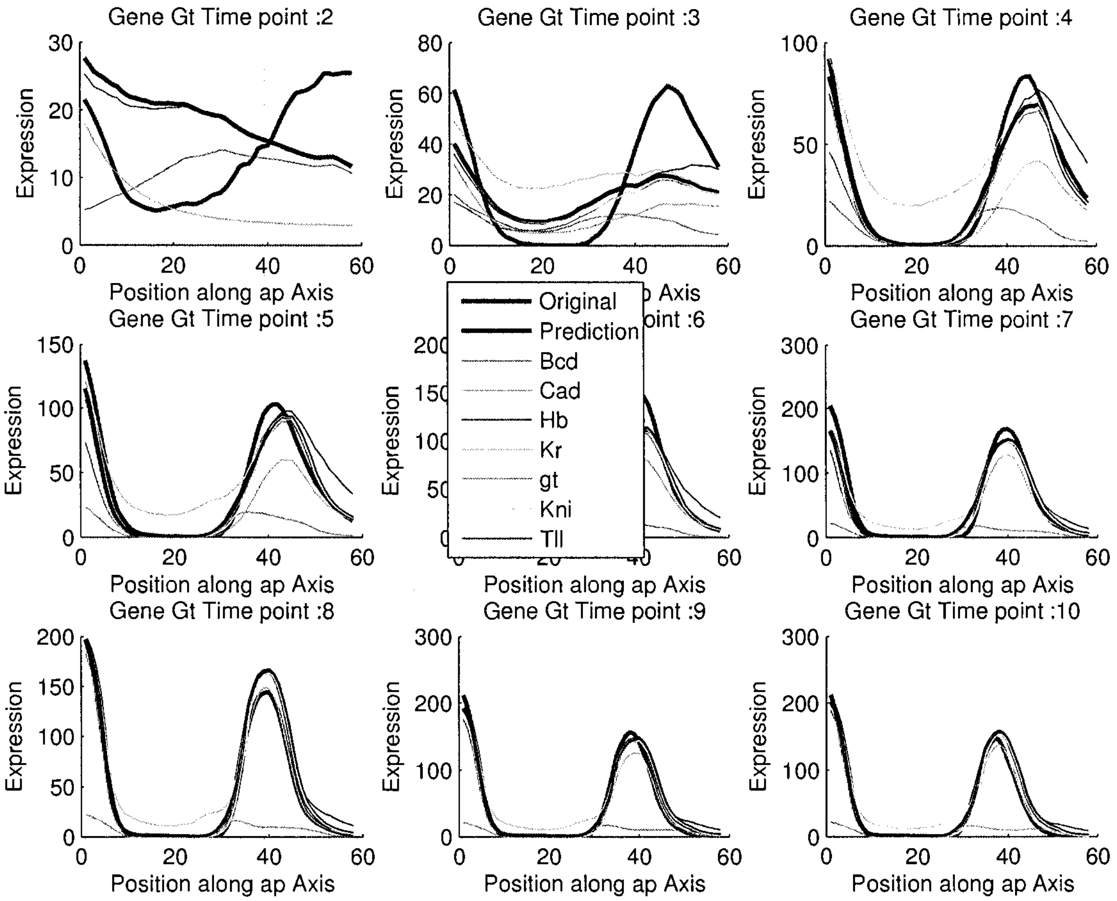


Figure 4–5: Transition-based optimization contribution plots for Gt. The thick lines denote the original and the final expression profiles. The thin lines represent the effect of the absence of each TFs.

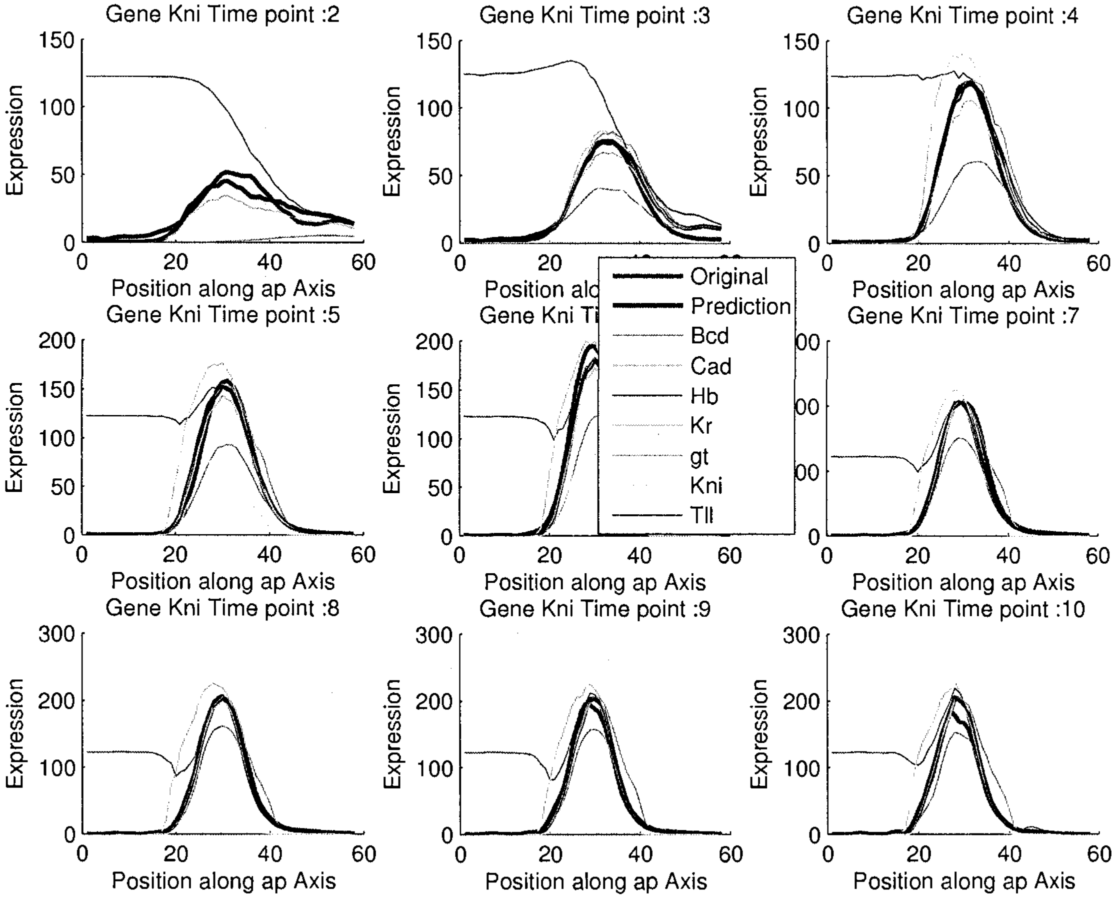


Figure 4–6: Transition-based optimization contribution plots for Kni. The thick lines denote the original and the final expression profiles. The thin lines represent the effect of the absence of each TFs.

I **Hunchback (Hb):** Figure 4-3 shows the absence plots for Hb as target for all the 9 step transitions. At time 2, the anterior Hb peak is formed due to autoactivation and Bcd activation. The posterior Hb is formed due to autoactivation only. However, this is not a valid mechanism, which exposes a potential flaw of this transition based modelling approach, i.e. the ability to explain the formation of peaks by autoactivation only. Knirps plays a part in forming the anterior edge of the posterior peak. The quality of prediction is usually excellent. The only feature that it misses is the final time point dip in the anterior *hb* domain, but this is a common problem with all the quantitative analysis studies in the literature. [42]

II **Kruppel (Kr):** The predictions at the earlier time stages are not very good. At later time steps, it gets better. The anterior border is influenced by Hb repression and the posterior border is influenced by Kni repression. Bcd acts as the principal activator for *Kr*.

III **Giant (Gt):** The predictions at the first two time steps are not very satisfactory. Like *Kr*, the predictions get better for the later time steps. Activation from Bcd and repression from Kr plays the key role in the formation of the anterior peak. Cad activation plays an important role for posterior peak construction.

IV **Knirps (Kni):** Figure 4-6 shows the absence plots for *kni* as target.The prediction performance is good for all the time points. Anterior border is formed by Hb and Kr repression. Bcd is the main activator for Knirps. Autoactivation also plays a role for the later time steps.

4.2.4 RMS Error

The overall RMS error is 10.46. The RMS errors for different target gene prediction at each time step is listed in Table 4-3 and in Figure 4-7. From Table 4-3,it is observed that *gt* has the greatest average error (12.0329) followed by *Kr*(11.1940), *hb*(9.4535) and *kni* (6.5446). The RMS error distribution for different time step is reported in Figure 4-7. A common phenomena observed for all the targets is that the final time point prediction quality is always worse than the immediate past time point. The error for *Hb* increases quite dramatically at the final time step. *gt* and *Kr* shows variations of error at the earlier time steps while *kni* maintains quite uniform low error predictions.

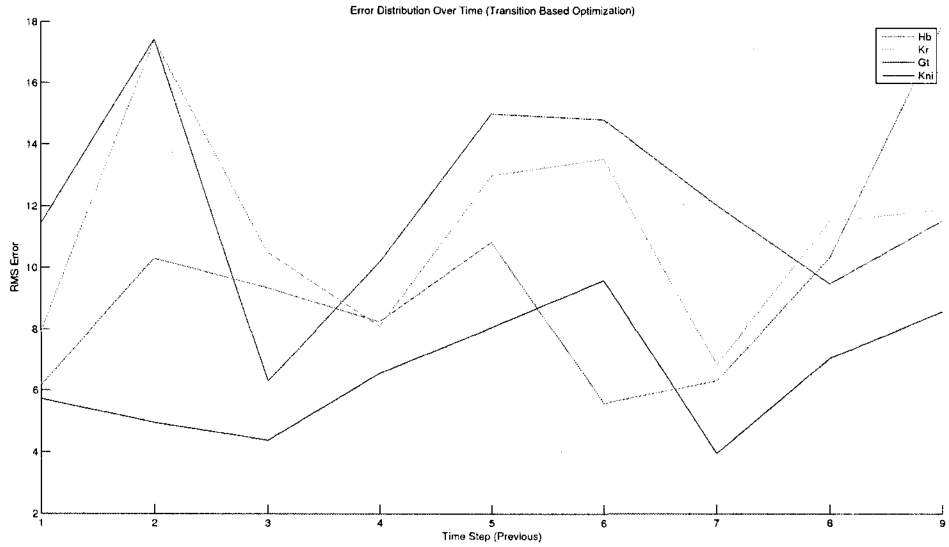


Figure 4–7: Transition-based optimization error distribution of different target over time

4.3 Trajectory-based Modelling Results

4.3.1 Model Fitting

We first attempted to fit the model exactly the same way we have done in Transition-based modelling. However we faced difficulties in finding a good fit using only simulated annealing due to the complexity of the model. We came up with a

Table 4–2: Other parameters obtained from the Transition-based fits

	θ Value	**R value**
Hb	0.2563	80.6788
Kr	1.2327	44.3763
Gt	0.1400	91.0843
Kni	1.2316	122.2731

Table 4–3: RMS errors for different target genes for all time steps(transition-based fits)

	Target Hb	**Target Kr**	**Target Gt**	**Target Kni**
Time Step 2	6.1737	7.9110	11.4280	5.7227
Time Step 3	10.3043	17.3890	17.4249	4.9515
Time Step 4	9.3511	10.5095	6.2980	4.3836
Time Step 5	8.2504	8.0968	10.1955	6.5633
Time Step 6	10.8473	13.0002	15.0081	8.0679
Time Step 7	5.5859	13.5464	14.8250	9.6027
Time Step 8	6.3048	6.8413	12.0608	3.9558
Time Step 9	10.3548	11.5444	9.4966	7.0607
Time Step 10	17.9089	11.9076	11.5592	8.5934

hybrid optimization strategy that incorporates both SA and Randomized local search for finding out a good fit for the model. The steps are discussed in details.

Initialization of parameters

The w parameters are initialized with very small random numbers. The initial signs of the weights are set to be the same as the unconstrained optimization fit results of [42] . The λ parameters are initialized to 0.5 and the R parameters are initialized to 105.0.

Randomized Local Search (RLS)

From the above initial estimates, we first numerically compute the gradient $\frac{\partial E}{\partial w_{ij}}$ for each i and j by calculating the effect of a small perturbation on the w_{ij}. A weight w_{ij} is then randomly chosen where the probability of choosing w_{ij} is proportional to the absolute value of $\frac{\partial E}{\partial w_{ij}}$. The chosen w_{ij} is then perturbed by taking a small step α_{ij} towards the direction opposite to the computed gradient $\frac{\partial E}{\partial w_{ij}}$ in the hope that it would improve the prediction. At the same time, small random perturbation to the λ_a and R_a parameters are performed. If the net effect of all the changes decreases E, we retain the changes. If not, we discard them.

The α_{ij} parameters are updated at each step. For this purpose, for each w_{ij} parameter, we keep the direction of the 'last good change', i.e. the last change that reduced E. If the current step is a 'good' one resulting in a lower error, and the current direction of change of w_{ij} is the same as the last good change of w_{ij}, we increase α_{ij} by 1%. If the direction of the current good change for w_{ij} is opposite to the last good change of w_{ij}, α_{ij} is lowered by 1 %. The opposite procedure, which decreases α_{ij} if the two directions are the same and increases α_{ij} if otherwise, is followed if the current step is a 'bad' one. However, if consecutively 100 updates are discarded, the α_{ij} parameters are reset to their initial value.

At each call to the randomized local search function, the initial estimates of the α_{ij} are passed as parameters and in total 5000 updates take place. We call this function repeatedly and decrease the initial estimates of the α_{ij} parameters if the number of successful updates (good changes) is less than 20. If the number of successful update is greater than 1000, we increase the initial estimates of α_{ij} . We repeatedly call this function 400 times which amounts to total 200,000 updates. After that we run the simulated annealing process.

Simulated Annealing

The output of the RLS algorithm serves as the input for the simulated annealing algorithm . We start with a higher temperature and gradually cool it until we get a reasonable solution and/or the parameters converge to a stable configuration.

4.3.2 Weights Obtained

The weights obtained are listed in Table 4.3. The signs of the weights generally conforms with the sign of unconstrained (unc_gc) fits as reported in [42] except that the signs for Cad and Tll on *hb* are swapped and the weight for Kni on *gt* is toggled. The Cad and Tll weight on *hb* is however supported by the Jaeger et al. [29] model. The relative magnitude of the weights significantly differ with the unc_gc fits. The cases where the magnitudes differ significantly with the unc_gc fits are listed below:

I *Cad* is now a strong activator for *hb*.

II Reduced activation of Bcd on *hb*, reduced repression of Kr and Gt on *hb*. Tll has a very small weight for *hb*. The contribution plots experiments will show

Table 4–4: Weight matrix obtained from trajectory-based Fits

	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.0137	0.2327	0.5858	0.1137
Cad	0.0123	0.0129	0.0260	0.0199
Hb	0.0370	-0.0855	-0.0896	-0.1111
Kr	-0.0052	0.0968	-0.2819	-0.0248
Gt	0.0045	-0.0964	0.0106	-0.0621
Kni	-0.1066	-0.1019	-0.0173	0.0447
Tll	-0.0031	-0.0125	-0.0039	-0.1680
Baseline	-3.5	-3.5	-3.5	-3.5

that these changes result in a different interpretation of the *hb* posterior peak formation.

III Increased effect of *gt* on *Kr,* reduced weight for *tll* on *Kr.*

IV Increased effect of *Bcd, hb, Kr and kni* on *gt.*

V Reduced effect of *gt* on *kni.*

4.3.3 Contribution Plots

By analyzing the *hb* absence plots (Figure 4-8), we observe that Bcd and Cad activates the anterior peak, Kni represses it in the middle and Cad activates the posterior peak. We also notice that auto-activation plays an important role .

For *Kr,* (Figure 4-9) the initial time step prediction is visually bad again.Bicoid activation is necessary for the formation of *Kr* peak. The precise positioning of the

Table 4–5: Other parameters obtained from Trajectory-based Fits

	θ Value	**R value**
Hb	-0.7310	131.3862
Kr	0.6177	76.8147
Gt	-3.1147	211.3760
Kni	-0.0892	116.2671

anterior edge of the peak is determined principally by *gt* and *hb* repression while the posterior edge is determined by *kni* repression.

For *gt*, the most dominant effect is of *Kr* repression. Bicoid and Caudal produce the anterior and the posterior peaks respectively. Hb and Kni repression is necessary for ensuring the right domain of expression for the posterior *gt* peak. (Figure 4-10)

Kni (Figure 4-11) is strongly repressed by Hb and Tll and activated by both Bcd and Cad which is conforming with the literature.

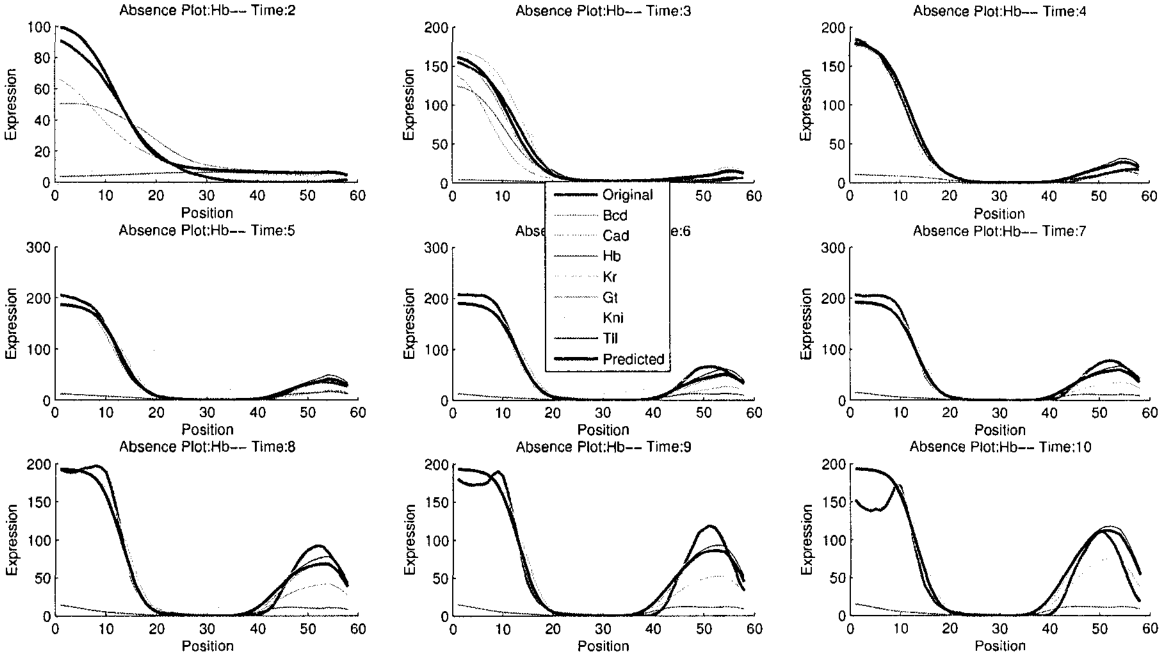


Figure 4–8: Trajectory-based optimization contribution plots for Hb. The thick lines denote the original and the predicted expression profile. The thin lines represent the effect of the absence of each TF.

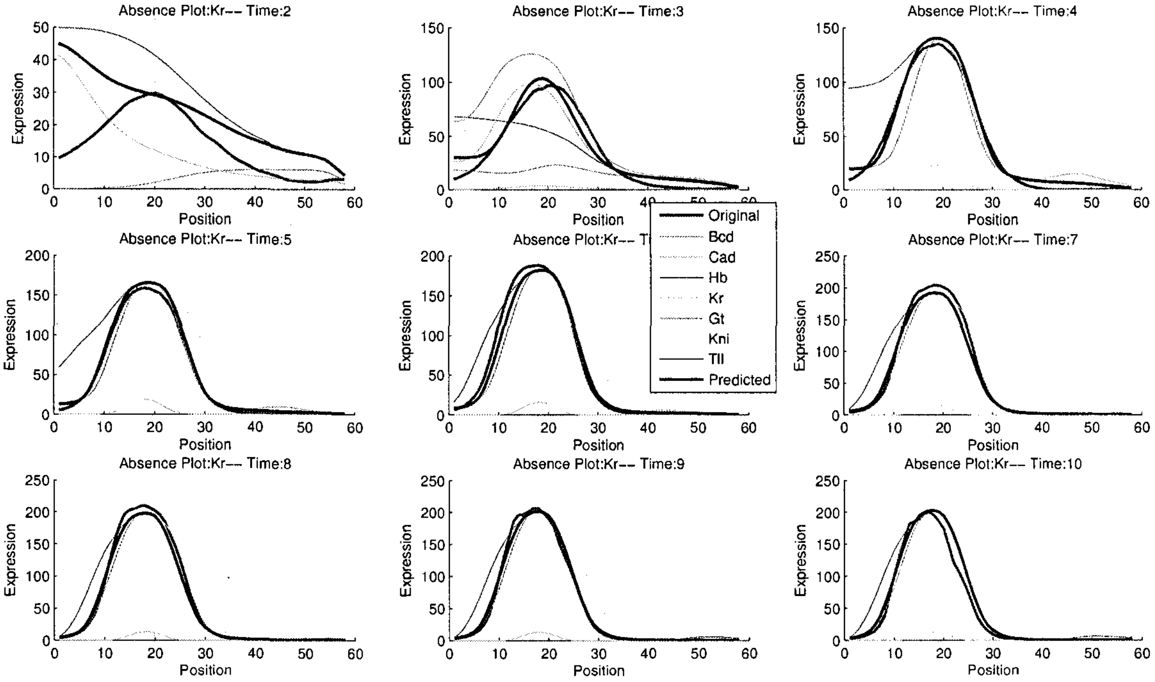


Figure 4–9: Trajectory-based optimization contribution plots for Kr. The thick lines denote the original and the predicted expression profile. The thin lines represent the effect of the absence of each TF.

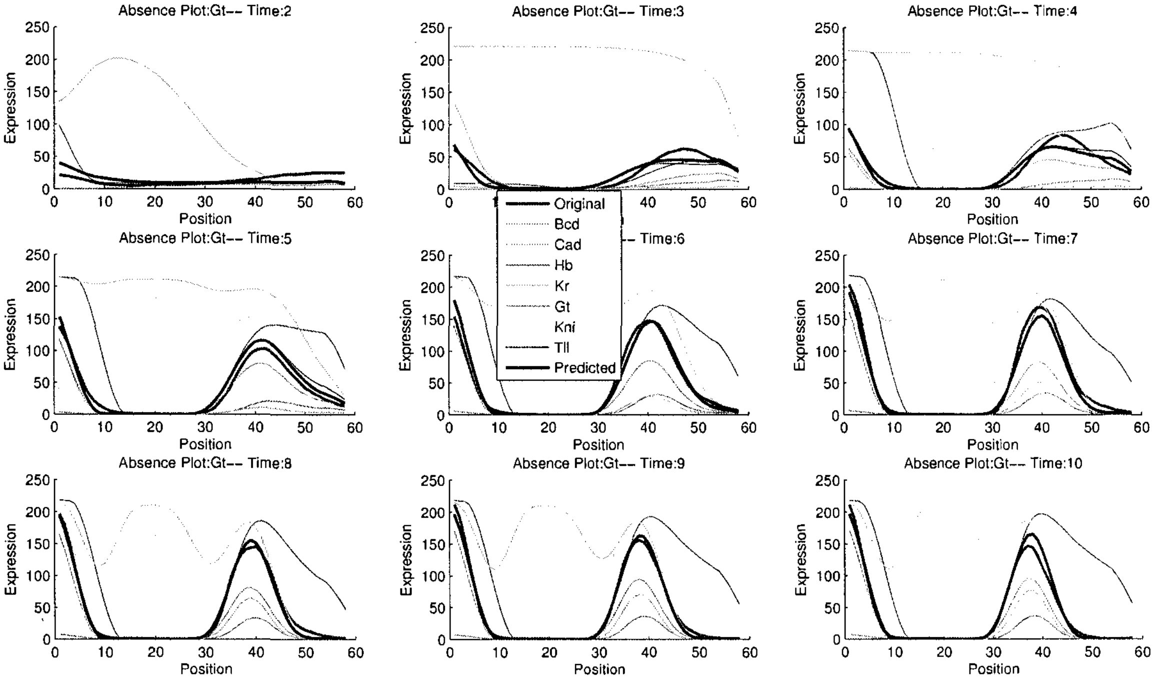


Figure 4–10: Trajectory-based optimization contribution plots for Gt. The thick lines denote the original and the predicted expression profile. The thin lines represent the effect of the absence of each TF.

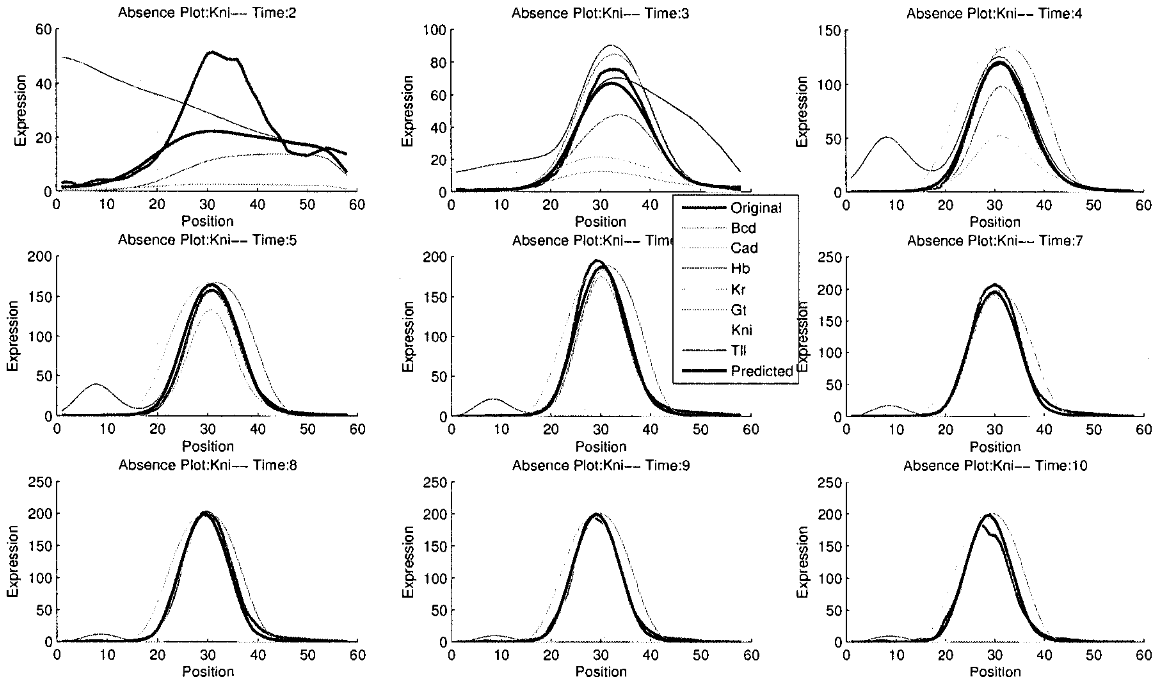


Figure 4–11: Trajectory-based optimization contribution plots for Kni. The thick lines denote the original and the predicted expression profile. The thin lines represent the effect of the absence of each TF.

4.3.4 Error Analysis

The overall RMS error is 9.0726. From Table 4-6, it is observed that *hb* has the greatest average error (10.3944) followed by *Kr*(8.5054), *gt*(8.0771) and *Kni* (6.4538). The rms error distribution for different time step is reported in Figure 4-12 It shows that for *gt, Kr* and *kni,* the error distribution is more or less uniform. But for *hb* the error increases for the later time stage . A common phenomena observed again for all the targets is that the final time point prediction quality is always worse than the immediate past time point.

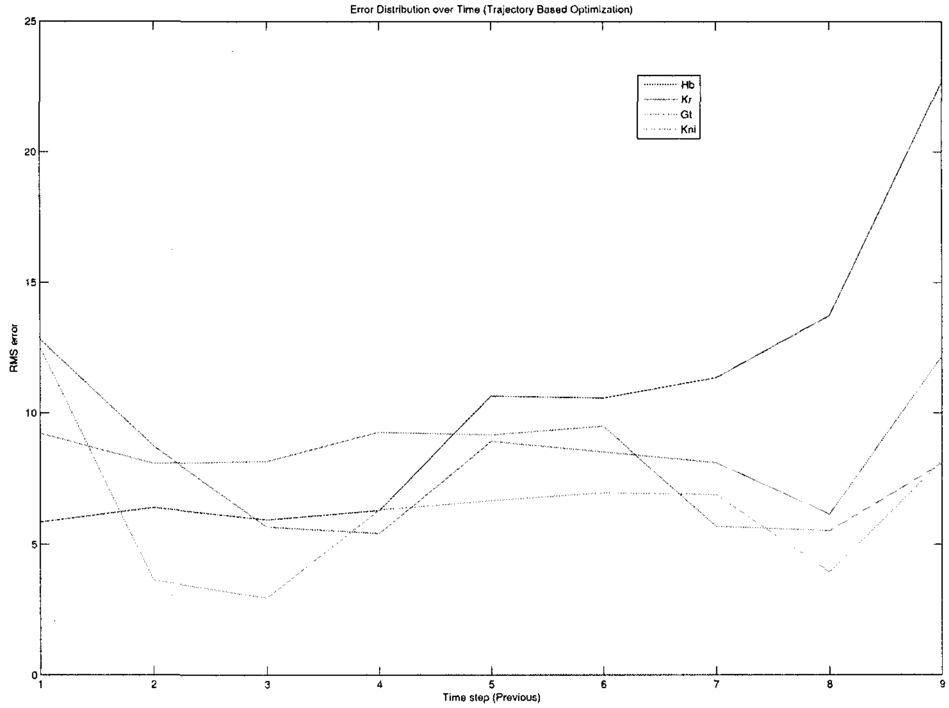


Figure 4–12: Trajectory-based optimization error distribution of different target over time

4.4 Comparison of Discrete-Time Models Results with Literature

In terms of prediction performance, discrete time gene circuit models exhibit less RMS error than its differential equation based models counterparts. The RMS error is only about 10 while both Jaeger and Perkins model yield more than 12. Despite the difference in the set of weights and parameters obtained and RMS error, there are significant amount of agreement among these models.

Table 4-7 compares how the different models attributes the key features to different TF activities. For anterior *hb*, Perkins et al. UNC GC fits attributes that to Bcd activation, Jaeger et al. to Bcd, Cad and autoactivation. Our trajectory-based model agrees with Jaeger et al. results. Transition-based model identifies Bcd and auto activation, but Cad activation is not a deciding factor for anterior *hb* formation. For posterior *hb*, both Jaeger et al. and trajectory Based model explains its formation with Cad activation. Perkins et al. emphasizes on Tll activation, which none of our model finds as a strong player for this case. The transition-based model attributes it principally to autoactivation of Hb.

For the *kr* anterior border, all the models find Hb repression as a reason. Besides, gt repression is also identified as a cause by the Jaeger model and our trajectory-based model. For the posterior Kr border, all the four models find Kni repression as the strongest deciding factor.

For Gt anterior peak, all the four models agree and attribute it to Bcd activation and kr repression. For the posterior peak, cad is identified as a key player by all the models. However, while Jaeger et al. and our transition-based result finds strong tll repression being one of the reasons, our trajectory-based model finds simultaneous inputs from Kr and Hb repression to be acting as a key contributor.

For Knirps, the anterior border formation is always attributed to Hb and Kr repression. The posterior border is attributed to gt. However, Jaeger model finds a significant input from Tll as well which is not supported by any other models.

4.5 Summary of Dynamical Models & Fits

Our deductions from the dynamic fit results are:

i Discrete-time models yield better predictive performance for the data set. Moreover the key factors for the formation of important features, as identified by these models, are generally in agreement with the existing literature.

ii Transition-Based model perform pretty well in terms of RMS error despite its simplicity. Their explanation of the formation of different features also agree to the literature and the trajectory-based models. The only exception is for posterior Hb model, it does not find significant contribution from Cad or Tll, which does not conform to the findings of the other models.

iii The complementary gap gene pair hb-kni and Kr-gt are always strongly mutually repressive, which is supported by all the other models. [42, 17, 36, 10, 25, 52, 12]

iv The only major problem that these models face is their inability to explain the gap gene shift with the proper cause. According to Jaeger et al, a chain of repressive interactions hb ⊣ gt ⊣ kni ⊣ kr causes the domain shifting of the gap genes, and their reverse interaction should not exist. However, the transition-based model does not find hb ⊣ gt repression. The gt ⊣ kni repression and kni ⊣ Kr repression is found but also kni ⊣ kr repression is detected. The trajectory-based model finds all these interactions, but it also detects kni ⊣ gt and kr ⊣ kni interactions which contradicts Jaeger et al results.

v For both the models, the expression of Knirps is the easiest one to fit. Also the final time point always exhibits more RMS error for all the target genes in both the models.

vi Both the models have detected auto-repression of the gap genes. However, the transition-based model finds a very small weight for Kni auto-repression.

In the next chapter, we attempt to relate the sequence-based data of binding site information with the expression-based regulatory weights obtained from our trajectory-based model. We have chosen the trajectory-based model for further investigation because from a biologist's viewpoint, this model is more realistic than the other models. Still, the model can account for the formation of the key features in the expression profile correctly enough and it also yields the greatest predictive performance in comparison with all the other models.

Table 4–6: RMS errors for different target genes for all time steps(Trajectory Based Fits)

	Target Hb	**Target Kr**	**Target Gt**	**Target Kni**
Time Step 2	5.8559	12.8212	9.2034	12.5222
Time Step 3	6.3998	8.7436	8.0882	3.6213
Time Step 4	5.9442	5.6641	8.1713	2.9505
Time Step 5	6.2973	5.4205	9.2622	6.3219
Time Step 6	10.6584	8.9183	9.1612	6.6819
Time Step 7	10.5868	8.5242	9.5042	6.9664
Time Step 8	11.3645	8.1222	5.6913	6.8982
Time Step 9	13.7384	6.1506	5.5225	3.9302
Time Step 10	22.7046	12.1844	8.0893	8.1913

Table 4–7: Comparison between models for the key contributors forming features. aa denotes auto-activation.

Features	Perkins et al.	Jaeger et al.	Transition-Based	Trajectory-based
A. Hb Peak	Bcd act.	aa+(Bcd +Cad)act.	aa+Bcd act	aa+ (Bcd.+Cad) act.
P. Hb Peak	Tll act	Cad Act.	aa.	Cad act.
Kr A. Border	hb rep.	(hb+gt)rep.	hb rep	(hb+gt) rep.
Kr P. Border	Kni rep.	Kni rep.	Kni rep.	Kni rep.
Gt A. Peak	Bcd act.+kr rep.	Bcd act.+kr rep.	Bcd act.+kr rep.	Bcd act.+kr rep.
Gt P. Peak	Cad act.	Cad act.+tll rep	Cad act.+tll rep	Bcd +Cad act.,(Kr+ Hb) rep.
Kni A. Border	Hb + Kr rep.	Hb + Kr rep.	Hb + Kr rep.	Hb + Kr rep.
Kni P. Border	Gt rep.	tll/gt rep.	Gt Rep	Gt rep.

CHAPTER 5
Associating the Binding Site Data with the Regulatory Weights

In this chapter, our attempt is to relate two different factors (sets of data),

i the regulatory weights obtained in our models, and

ii the binding of the transcription factors with their respective binding sites.

Unfortunately, there is no perfect universal model for these factors. Different algorithms exist for counting the occupancy or strength of the binding sites present in the regulatory region of a target gene. We can obtain two totally different sets of cumulative binding site strength data by using two different algorithms. In the same manner, different models yielded different sets of data as reported in chapter 3,4 and the relevant previous works [42, 29, 50, 53]. None of the models for TFBS finding and regulatory weight determination can fully replicate all the underlying biological processes. Moreover, there is no convenient way of cross-validating the results. As a result,we cannot be sure if the values obtained by any of the models are actually representing the true values of the regulatory weights and the binding site strengths.

However, from a high level viewpoint, the binding site strengths are the "causes" and the regulatory weights are the "effects" and the true numerical values of these two factors must be related through a function. The function that describes this relationship is unknown to us. As a starting point, we can assume a proportional relationship between them. In this chapter, our aim is to estimate the true value of these two important factors, i.e. we would like to find a set of regulatory weights that can explain the dynamics of the expression and at the same time, can retain the relationship of proportionality with the corresponding CBSS matrix elements and vice versa. With this end in view, we analyze the effect of updating one set of data (either the expression-based or the sequence-based estimates) at a time according to an assumption of proportionality between the two sets of data, keeping the other data set fixed. Before we move on to discuss the results of such updates, we report how our obtained weights from the trajectory-based optimization relate to the cumulative binding site strength data.

5.1 Binding Site Analysis for Trajectory-based Optimization

The binding site data that we have used in Chapter 3 were taken from Schroeder et al. [55] results. However, the binding site information of Schroeder et al. seems to have some missing values, especially for Knirps CRMs. Exceptionally high values of BS strengths were observed for Tailless as a TF. For this reason, we opted for using the algorithm Stubb [59, 58]. for binding site calculation instead. We ran the algorithm using the PWMs listed in Sinha et al. on the CRMs identified by the Schroeder paper. We also changed the prior probability parameter of stubb to 0.001 instead of the standard value of 0.01 because we observed that the prior probability value of 0.01 tends to give many false positives when we randomly permute the nucleotides of the CRM sequence. Our studies had shown that the binding site strength output using a prior probability value of 0.001 approximately matches the binding site strengths obtained by using prior probability 0.01 with false positive subtraction. As false positive estimation requires randomization of the sequence,

Table 5–1: Cumulative binding site strength matrix output of Stubb

	Hb	**Kr**	**Gt**	**Kni**
Bcd	2.7870	4.4184	5.4254	3.6769
Cad	1.1289	2.4806	4.1644	1.8496
Hb	2.7051	5.4949	5.0298	4.7593
Kr	4.1529	2.5365	10.5966	4.1236
Gt	0.5657	2.7116	1.6721	0.2670
Kni	1.8683	2.1446	3.8105	2.4646
Tll	2.0230	2.4204	6.1302	3.2852

we avoided the use of such a time consuming procedure by directly using a prior probability value of 0.001. The CBSS matrix obtained is reported in Table 5-1.

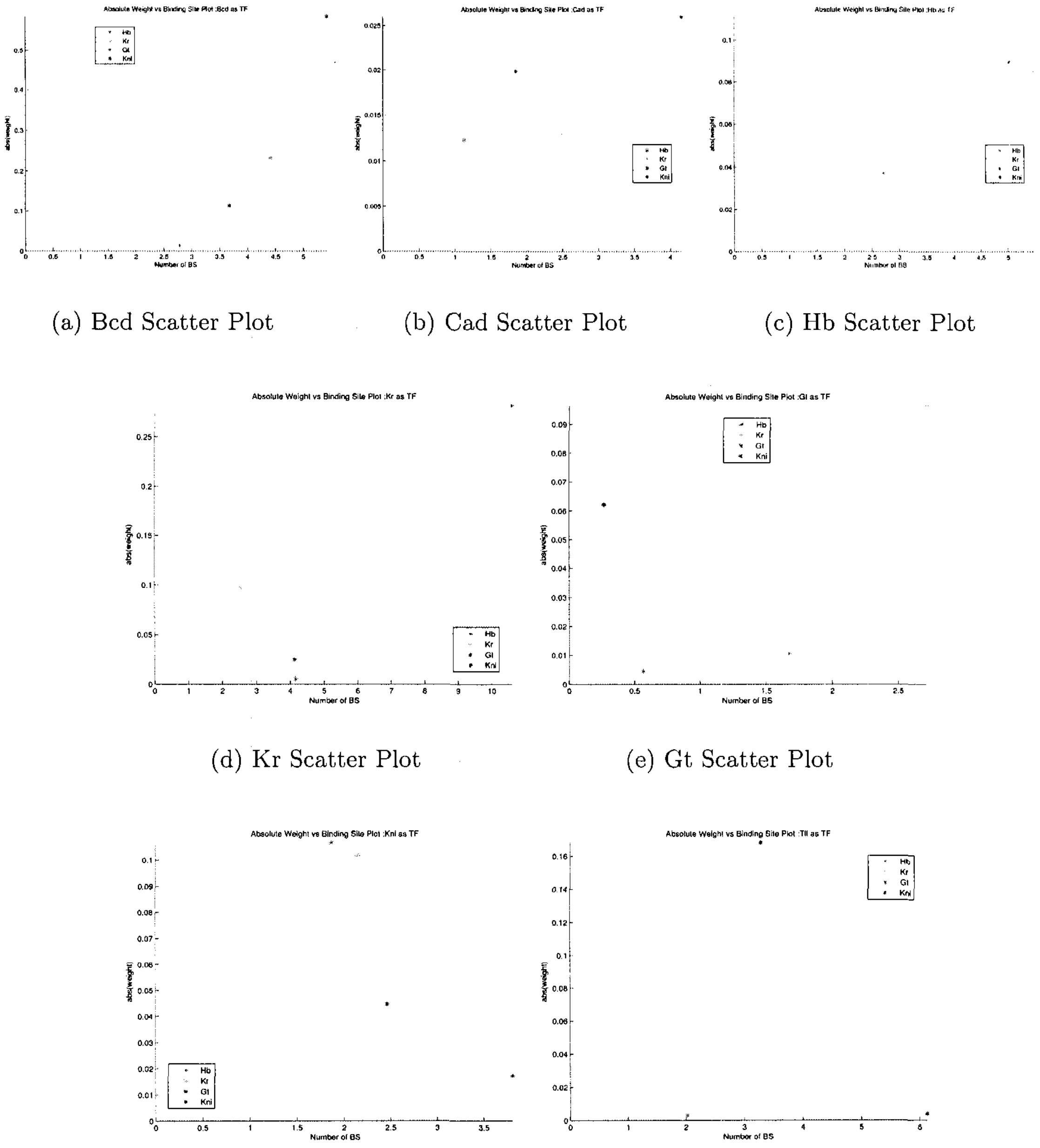
We have plotted the scatter plots of the binding site vs. regulatory weight magnitudes per TF (rows of the matrices) to visualize the relationship between these two factors. Note that only the magnitude (absolute value) of the regulatory weights are considered, not their signs. As TFBS occupancy, presented by the CBSS matrix, is the reason behind the regulatory effect of a TF on its targets, and the regulatory weight is a measure of the strength of that regulatory effect, we expect that the CBSS matrix and the regulatory weights should be correlated. If we assume that the relationship between them is linear, correlation may be a good measure of the dependency among these two sets of data. We measure the per TF correlation between these two matrices. Per TF implies a fixed *TFBS potency* assumption, i.e. per TF correlation of +1 implies that if a new binding site for a given TF at the regulatory region of a target is occupied, it always increase the regulatory weight of that TF on the corresponding target by a fixed amount irrespective of the target gene in whose regulatory region the binding site is added.

Figure 5.1(a) shows that for Bcd as TF, the number of binding sites increases monotonically if the magnitude of weight increases. The correlation coefficient between the binding site strength data with the corresponding regulatory weight is 0.96. For Cad, Hb and Kr as TF, the correlation coefficient is about 0.80. But for Gt as TF, the correlation coefficient drops to about 0.50. For Kni as TF, we get surprisingly strong negative correlation (about -0.9). For Tll as TF, the correlation coefficient is nearly zero. From the plots and the correlation coefficients, the trend is towards a strong positive correlation for Bcd, Cad, Hb and Kr as TFs, weak positive correlation for Gt, negative correlation for Kni as TF and no correlation for Tll as a TF.

As the binding site data and regulatory weights were derived independently of each other, the majority trend towards a positive correlation may suggest that there exists a positively correlated relationship between the true value of binding site strengths and the true value of the regulatory weights. However, we also have to realize that correlation coefficient calculated from only 4 data points (one per target) may not be reliable enough a measure for the linear dependence. Besides, from a biological viewpoint, using correlation coefficient as a measure of linear dependence raises some serious concerns. In the next section, we describe the problem of using correlation coefficient in this case and then describe a suitable error criteria which measures not only linearity, but also proportionality between the two factors.

5.2 Measure of linear dependence

Although correlation coefficient is a widely employed metric as a measure of linear dependence between an independent and a dependent variable, it cannot be



(a) Bcd Scatter Plot

(b) Cad Scatter Plot

(c) Hb Scatter Plot

(d) Kr Scatter Plot

(e) Gt Scatter Plot

(f) Kni Scatter Plot

(g) Tll Scatter Plot

Figure 5–1: Binding site strength vs regulatory weights for the trajectory-based optimization results. X and Y axis denote Number of BS and absolute value of weight respectively. The scatter points represent different target gene for a fixed TF.

used as a measure of proportionality between these variables. If x is the independent variable and y is a dependent variable, a correlation coefficient of $+1$ implies that the rate of change of y w.r.t x is fixed, i.e. dy is proportional to dx, but that does not necessarily imply that y itself is proportional to x.

Mathematically, if we assume that for each TF b, the absolute value of the corresponding weights $|w_{ba}|$ and CBSS values $B(b,a)$ for all targets a should show a strong positive correlation, then in the perfect case, $|w_{ba}| = c_b \cdot B(b,a) + \alpha_b$ where c_b is the slope and α_b is the y intersect of the line that goes through all the points when we make similar scatter plots as Fig 5-1 having w parameter on the y axis and B parameters on the x axis. If we plug the value for w_{ba} into the model equation of trajectory-based models, the model equation becomes,

$$y_a^i(t+1) = D\left(R_a \cdot g\left(\sum_b (\alpha_b + c_b \cdot B(b,a) \cdot sign(w_{ba}) \cdot y_b^i(t) + h_a \right) + (1 - \lambda_a) \cdot y_a^i(t) \right)$$

which implies that if no binding site for TF b is present in the regulatory regionof target a (i.e. $B(b,a) = 0$), the production term still does not go to zero. Instead we get a term which is proportional to the expression profile of the TFs. This is contradictory to the fact that the attachment of TFs to their respective binding sites is the cause for transcription process initiation. Therefore it becomes clear that there should be no such α_b terms in this case.

Following this line of thought, when we impose a *proportionality constraint* on the definition of correlation assuming $|w_{ba}| = c_b \cdot B(b,a)$,i.e. $\alpha_b = 0$, then plugging the w_{ba} value into the model equation infers that if there is no binding site for a particular TF, it should not contribute at all towards the production of the target gene.

From the above discussion, it becomes apparent that the correlation coefficient is not a good metric for the measure of correlation in this particular case. What we actually need here is a measure of proportionality which will measure by what degree, the data satisfies the assumption that for each TF b, $|w_{ba}| = c_b \cdot B(b,a) \forall_a$. We devise an error term that is zero when this assumption is fully satisfied and which penalizes the data sets on the basis of their divergence from this proportionality assumption. We call this error term as E_{prop} where,

$$E_{prop} = \sum_b min_{c_b} \sum_a (|w_{ba}| - c_b * B(b,a))^2 \tag{5.1}$$

For all targets a, if $|w_{ba}| = c_b * B(b,a)$, then E_{prop} is zero. The larger the difference between $|w_{ba}|$ and $c_b * B(b,a)$ the larger the value for the inner sum. Let us term the inner sum result as ξ_b, i.e. $\xi_b(c_b) = \sum_a (|w_{ba}| - c_b * B(b,a))^2$. In order to find the minima of ξ_b with respect to c_b we equate the gradient of ξ_b to zero.

$$\begin{aligned}
&\frac{\partial \xi_b}{\partial c_b} = 0 \\
&\Rightarrow -2 \cdot \sum_a (|w_{ba}| - c_b \cdot B(b,a)) \cdot B(b,a) = 0 \\
&\Rightarrow c_b = \frac{\sum_a |w_{ba}| \cdot B(b,a)}{\sum_a (B(b,a))^2}
\end{aligned}$$

Plugging in the calue in Equation 5.1, we get

$$E_{prop} = \sum_b \left(\sum_a \left(\frac{\sum_{a'} |w_{ba'}| \cdot B(b,a')}{\sum_{a'} (B(b,a'))^2} \cdot B(b,a) - |w_{ba}| \right)^2 \right) \tag{5.2}$$

We use this definition of E_{prop} as an error metric for the lack of proportionality between $|w_{ba}|$ and $B(b,a)|$ in the subsequent sections.

5.3 Optimizing the regulatory weights to match the binding site data

In this section, we describe an optimization strategy for finding a regulatory weight matrix that can explain the regulatory dynamics of the system and is still able to retain high correlation (with a proportionality constraint) with the given CBSS values. We have used the negative of E_{prop} as in Equation 5.2 as a measure of correlation. The natural way to incorporate both these requirements is to augment the error function to be minimized by our fitting procedure to include both a prediction error component and a correlation component .

5.3.1 Augmenting the Error Criteria

We use the following augmented error function:

$$E \ = \ \beta \cdot E_{rms} + \gamma \cdot (1-\beta) \cdot E_{prop}$$

where,

$E_{rms} = \sqrt{\frac{\sum_{\forall_{i,a,t}} (y_a^i(t) - v_a^i(t))^2}{N}}$

$y_a^i(t)$ =Model Predicted Expression Level of gene a at position i in time t.

N= Total number of predicted samples. In this case, N= 58*4*9 as 9 time steps are predicted starting from the initial condition for 58 expression values each for 4 target genes *hb*, *Kr*, *gt* and *kni*.

γ = a multiplication factor for compensating for the difference of scale of E_{rms} and E_{prop}.

The $\beta \in [0, 1]$ parameter is a knob to tune the relative emphasis of the two error components. If $\beta = 0$ then the augmented error function contains only the component with E_{prop}. When the proportionality constraint is perfectly satisfied, the line joining the points on the scatter plot goes through the origin and $E_{prop} = 0$. For any other case, $E_{prop} > 0$, which will increase the augmented error.

As β grows from 0 towards 1, more and more emphasis is given on the accuracy of the prediction by sacrificing the proportionality constraint. If $\beta = 1$ then the augmented error function contains only the E_{rms} component. In this case, the model fitting procedure will concentrate on an accurate prediction only, not caring about the proportionality at all.

We use β values from 0.1 to 1.0 with step size 0.1 to get 10 different sets of optimal parameters. Note that the $\beta = 1$ case is the unconstrained optimization (trajectory-based) that minimizes the RMS error only. The value of γ is set to 2000.

We used the same optimization procedure of running SA on the randomized local search result as we had done during the trajectory-based model optimization.

5.3.2 RMS and Correlation Values for Different Values of β

Table 5-2 lists the RMS error and E_{prop} values obtained from the different fits with varying β. As expected, as we increase the value of β, the RMS error decreases and (E_{prop} increases. From our analysis on the numerical values of E_{prop}, we found out when $E_{prop} > 10^{-5}$, the visual inspection of the scatter plots (per TF) reveals that the points on the many of the plots do not form a line going through the origin. From our results, we observe that only $\beta = 0.1$ case yields a E_{prop} value which is less than 10^{-5}. In order to gain insight of the results, we choose this particular case for analysis in details.

Table 5–2: RMS error and E_{prop} values for different values of β

	E_{rms}	E_{prop}
0.1	23.62	6.19 x 10^{-5}
0.2	22.30	1.95 x 10^{-4}
0.3	21.36	3.61 x 10^{-4}
0.4	20.61	5.74 x 10^{-4}
0.5	19.92	8.88 x 10^{-4}
0.6	18.90	0.0015
0.7	17.14	0.0031
0.8	14.74	0.006
0.9	13.17	0.011
1.0	9.07	0.156

5.3.3 Weights obtained when $\beta = 0.1$

The weights obtained are listed in Table 5-3 and the other parameter values are reported in Table 5-4. We observe that the signs and the magnitude of the weights differ greatly with the weights obtained from the previous models. Bicoid has been identified as a repressor for Giant and Knirps, which is unprecedented in any of the models seen so far. Knirps is termed as an activator for Hunchback and Giant, while all the models identify Knirps as a strong repressor for Hb. The relative magnitude of many weights differ significantly with the other model results.

5.3.4 Contribution Plots

Although the RMS error is greater than 20, we have found that all the peaks were predicted by the model. However, the peaks are either shifted or have a greater width than the observed peak which was the reason behind for a greater RMS error.

Table 5–3: Weight matrix obtained when $\beta = 0.1$

	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.0953	0.1525	-0.1866	-0.1268
Cad	-0.0098	0.0166	0.0263	0.0157
Hb	0.0266	-0.0553	0.0534	-0.0488
Kr	-0.0189	0.0121	-0.0476	0.0135
Gt	-0.0074	-0.0345	0.0211	-0.0042
Kni	0.0125	-0.0165	0.0258	0.0204
Tll	0.0282	-0.0340	-0.0853	-0.0462
Baseline	-3.5	-3.5	-3.5	-3.5

By analyzing the *hb* absence plots (Figure 5-2), we observe that Bcd and auto activation activate the anterior peak, Kni represses it, Cad represses *hb* in the middle and Tll activates the posterior peak. We also notice that the posterior peak is shifted.

For *Kr,* (Figure 5-3) the initial time step prediction is visually bad again. Bicoid and Caudal activation is necessary for the formation of *Kr* anterior peak. The precise positioning of the anterior edge of the peak is determined principally by Hb repression while the posterior edge is determined by Kni and Gt repression. The peak is wider at the posterior end.

For *gt*, the most dominant effect is of Bcd repression. Caudal produces the anterior and the posterior peaks. Giant act as an activator for the posterior peak. Kr is the most significant repressor for the anterior peak (FIgure 5-4).

Table 5–4: Other parameters obtained when $\beta = 0.1$

	θ Value	**R value**
Hb	0.2477	97.0420
Kr	1.3470	67.7648
Gt	-1.1539	132.7759
Kni	-0.1449	177.4841

Kni (Figure 5-5) has a predicted domain much wider than the original expression. Bicoid is identified as a repressor. The posterior edge has been formed due to Tll repression and joint activation effect from Kr and Cad.

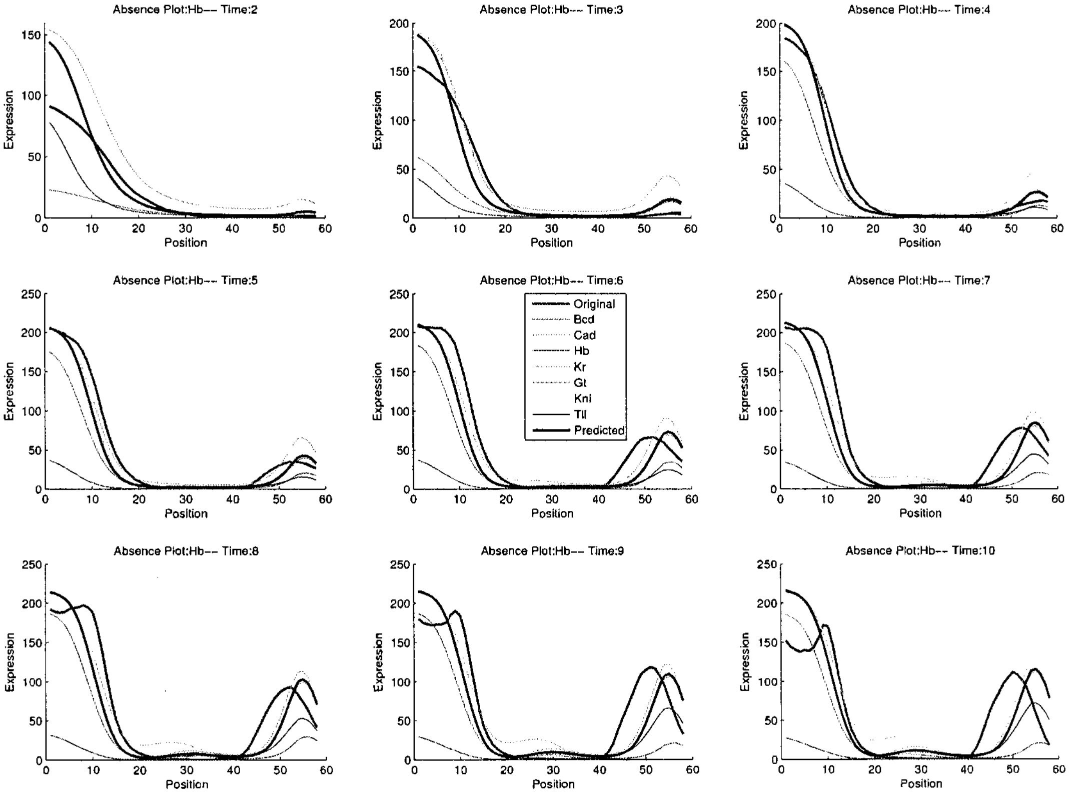


Figure 5-2: Contribution plots for Hb($\beta = 0.1$)

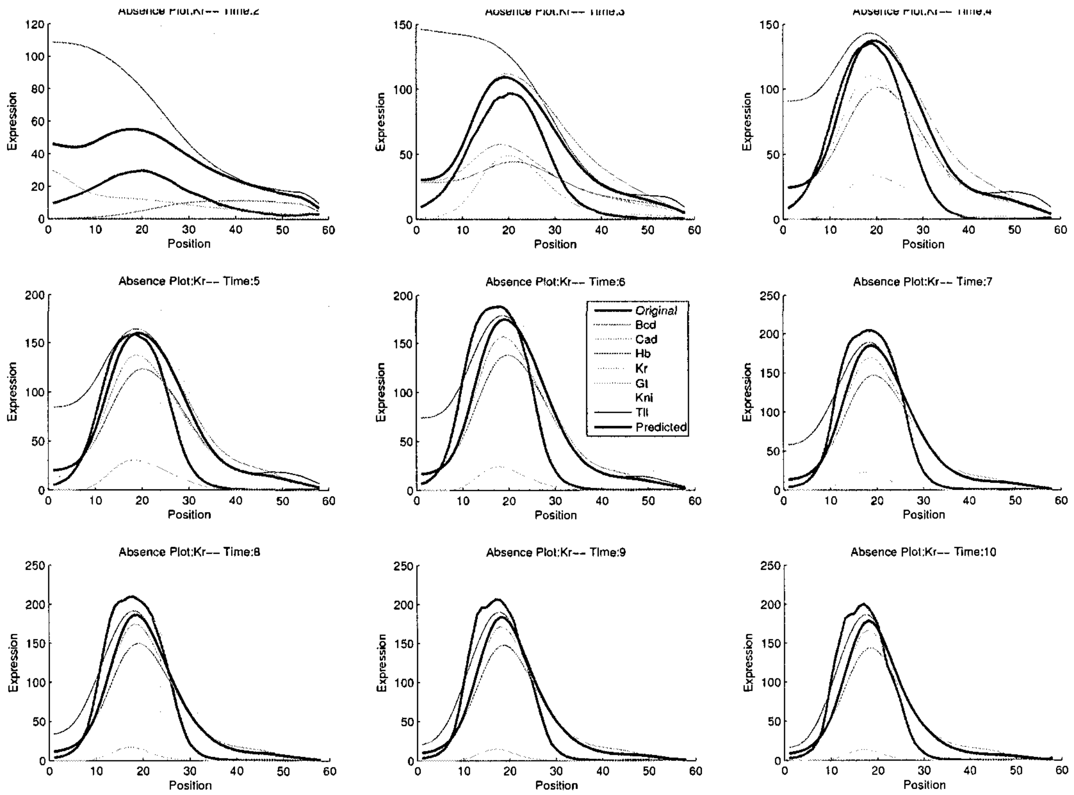


Figure 5–3: Contribution plots for Kr ($\beta = 0.1$)

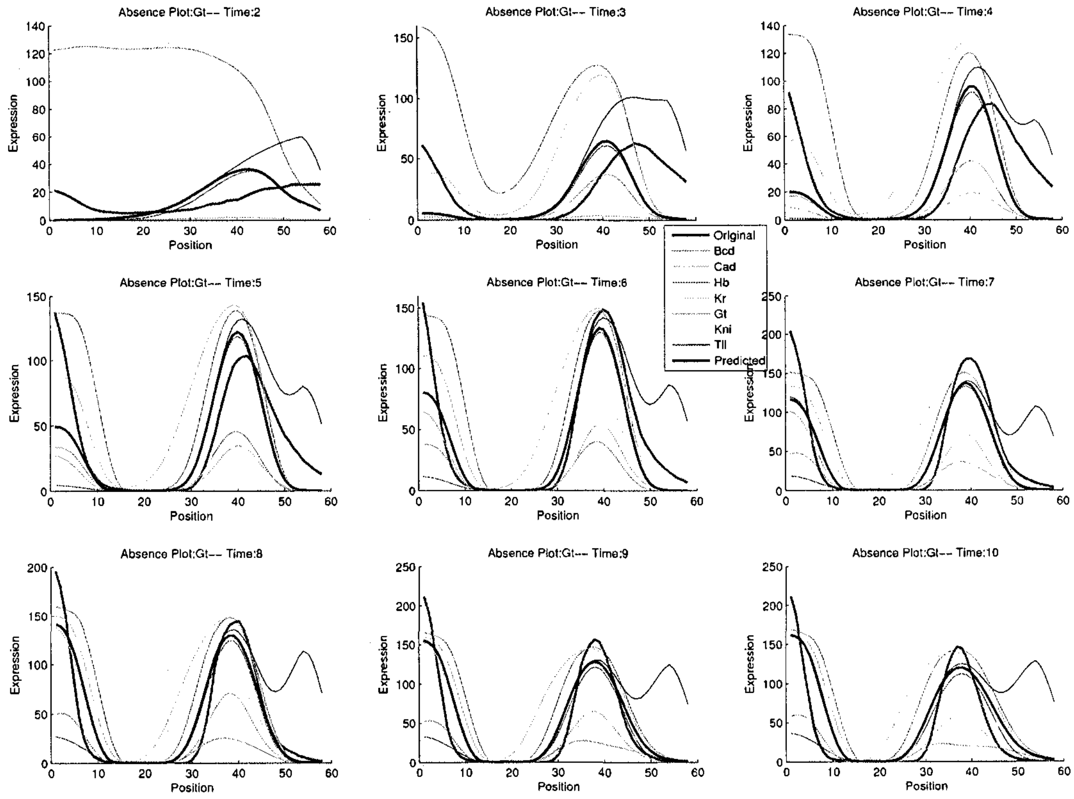


Figure 5–4: Contribution plots for Gt($\beta = 0.1$)

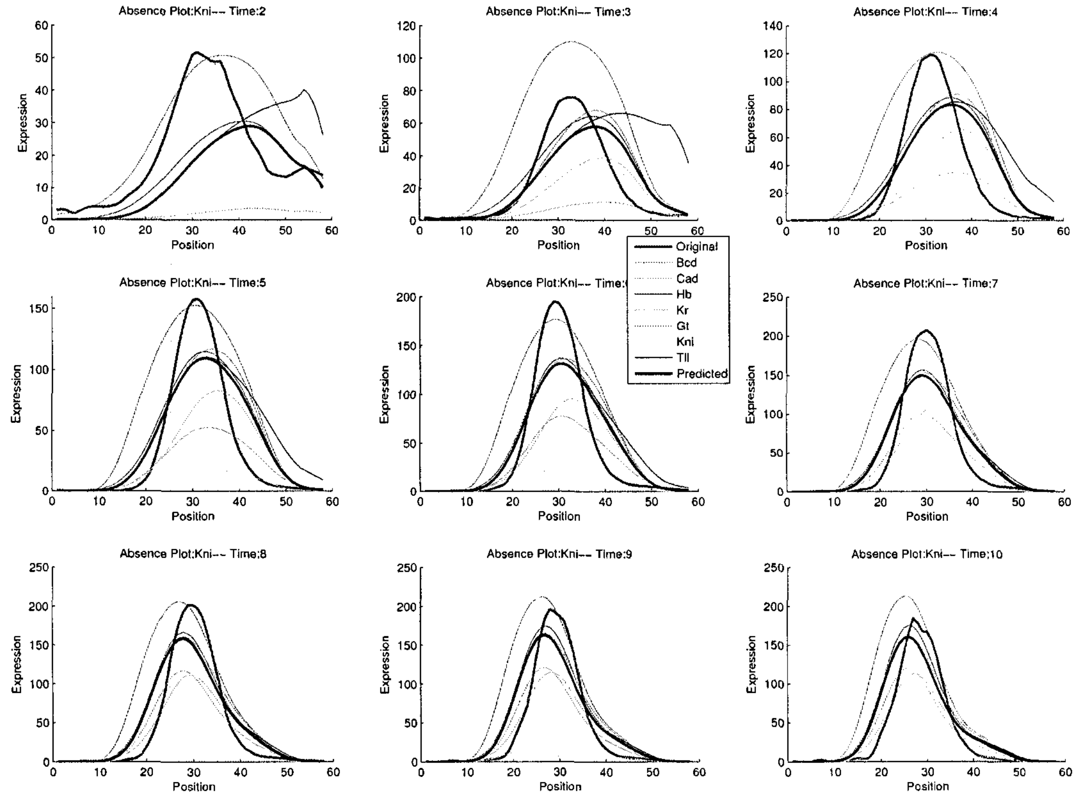
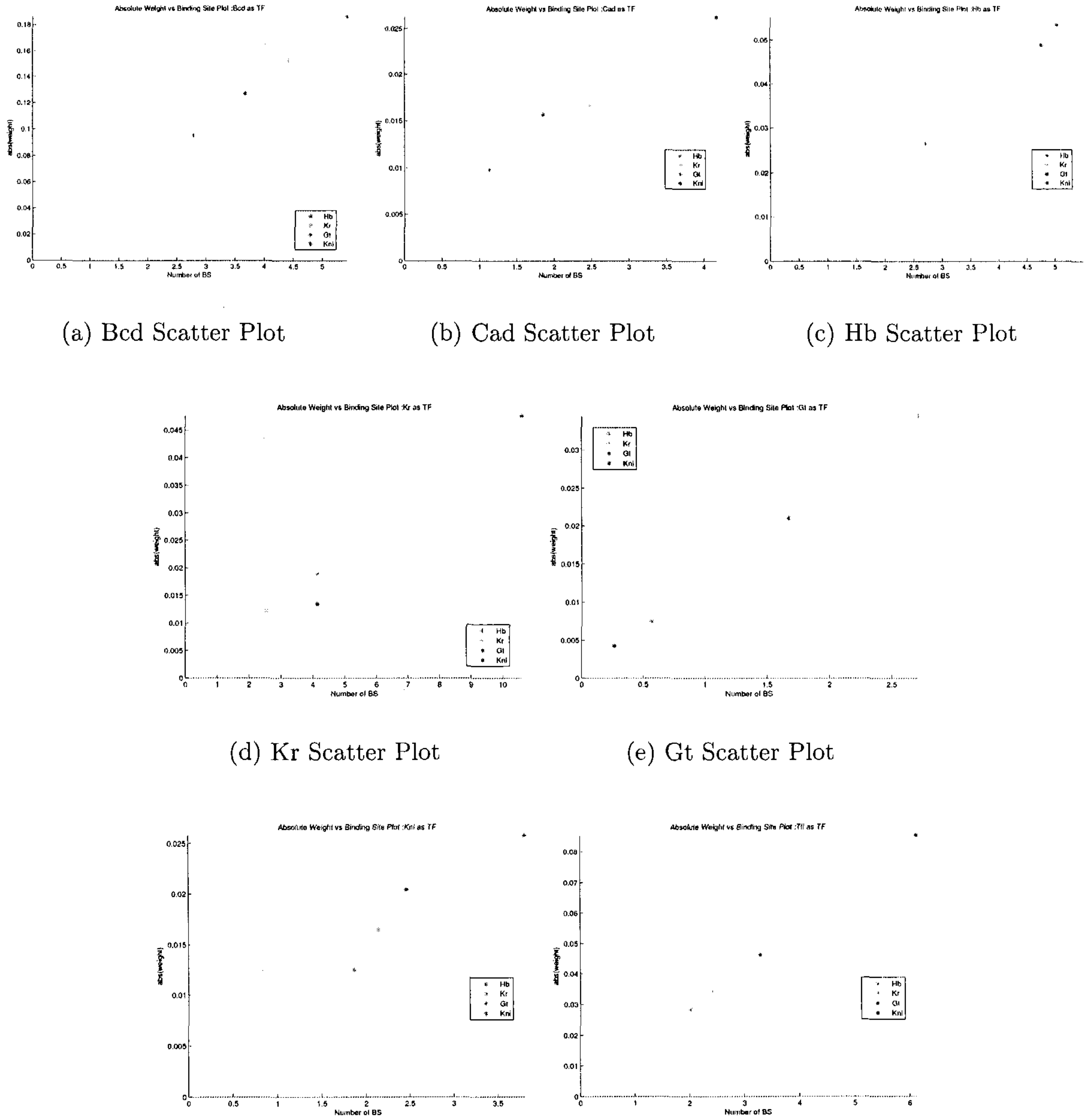


Figure 5–5: Contribution plots for Kni($\beta = 0.1$)

5.3.5 Correlation after optimization

Figure 5-6 presents the scatter plots after the optimization. We observe that the points on the scatter plot now form a line going through the origin for most of of the TFs. As β has a small value, more emphasis was given on maximizing the E_{prop} than the accuracy of prediction. As a result, the optimization result shows greater level of proportionality at the cost of a greater RMS error.



(a) Bcd Scatter Plot

(b) Cad Scatter Plot

(c) Hb Scatter Plot

(d) Kr Scatter Plot

(e) Gt Scatter Plot

(f) Kni Scatter Plot

(g) Tll Scatter Plot

Figure 5–6: Binding site strength vs regulatory weights when $\beta = 0.1$. X and Y axis denote Number of BS and absolute value of weight respectively. The scatter points represent different target gene for a fixed TF.

5.3.6 Error Analysis

The overall RMS error is 23.624. If we calculate the mean RMS error from Table 5-5, we will see that the mean RMS error lies within 22-24 for all the target genes. So the mean error is quite uniform for all the targets. The RMS error distribution for different time step is reported in Table 5-5 and Figure 5-7. It shows that the prediction performance deteriorates for the later time steps.

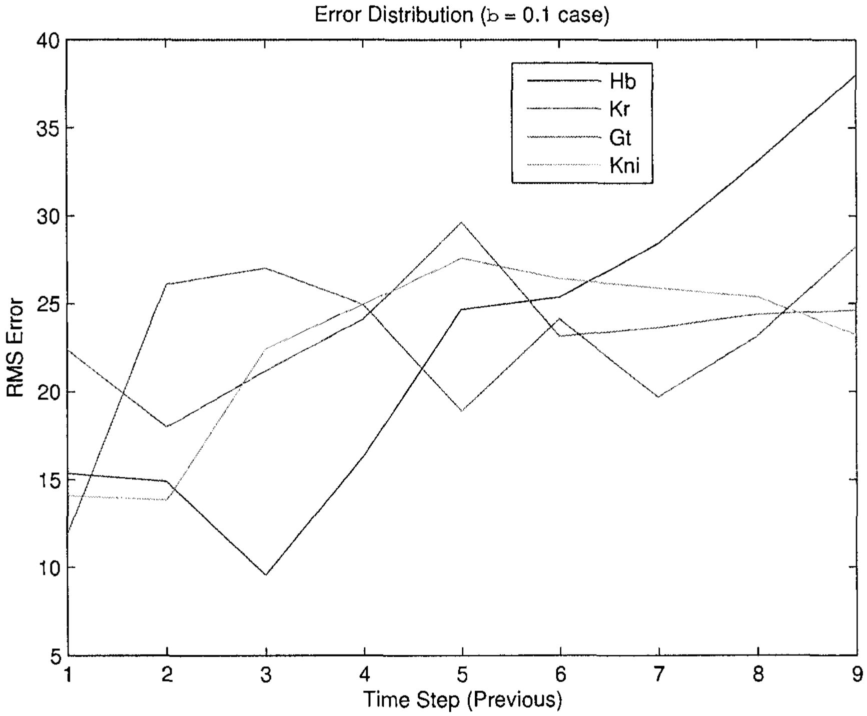


Figure 5-7: Error distribution of different target over time ($\beta = 0.1$)

5.3.7 Comparison of $\beta = 0.1$ Results with Literature and Trajectory-based Fits

In terms of prediction performance, the RMS error obtained is much higher than any of the models. The weights obtained also differ significantly from the other

Table 5–5: RMS errors for different target genes for all time steps($\beta = 0.1$ Fits)

	Target Hb	**Target Kr**	**Target Gt**	**Target Kni**
Time Step 2	15.3613	22.3930	11.8596	14.0780
Time Step 3	14.8795	17.9808	26.0895	13.8069
Time Step 4	9.5472	21.1865	27.0216	22.4129
Time Step 5	16.3408	24.1049	24.9239	24.9498
Time Step 6	24.6601	29.6334	18.8921	27.5756
Time Step 7	25.3580	23.1281	24.1120	26.3955
Time Step 8	28.4155	23.6231	19.6856	25.8742
Time Step 9	33.0636	24.3659	23.1231	25.3487
Time Step 10	37.9993	24.6129	28.2777	23.2063

models. The most significant difference is the identification of Bcd as a repressor for Gt and Kni which made the model weak. The gap genes have more often been termed as activators. Table 5-6 compares how the different models attributes the key features to different TF activities. The notable differences for this model are:

I Cad is a significant repressor for *hb*. The posterior *hb* is activated by Tll which agrees with Perkins et al. results.

II Bcd is a significant repressor for Giant and Knirps. Knirps is an activator for Hunchback and Giant.

III Tll repression plays an important role for the positioning of the posterior border of Knirps.

IV The predicted expression domains are usually wider than the original domains.

In brief, the optimization results differ significantly in its interpretation of the regulatory interactions and their effects. The prediction error is much larger. We still get all the peaks but their domain of expression does not precisely match with the original expression profile. This model does not find gt $\dashv$ kni repression, an interaction which was identified by Jaeger et al. as important for the gap gene domain shift. The opposite interactions of the domain shift hypothesis (the effect of Kr on *kni*, Kni on *gt* and Gt on *hb*) are usually activating while according to Jaeger et al., these reverse interactions should be neutral, i.e. no effect. Perhaps this is the reason why we get a wider domain of expression profiles.

To conclude, our attempts to optimize the weights to find a good correlation component with the binding site data fail to fit the data well. The regulatory interactions are also not identified properly. In the next section, we present the reverse optimization problem of optimizing the PWMs to change the binding site data such that the new binding site composition shows a higher level of correlation with the trajectory-based model weights.

5.4 PWM Optimization to Match the Regulatory Weights

The binding site strength data we have used in this chapter was obtained from the output of the Stubb algorithm [58] providing the position weight matrices given in [58] and the predicted CRMs of the Ahab Algorithm [55] as inputs. The PWMs were constructed from a set of known binding site motifs. However, for different TFs, the number of known binding site differs significantly. For example, Hb PWM was constructed from 43 known binding sites, while the Gt PWM was constructed from only 8 known binding sites [45, 55, 58] . Moreover, not all binding sites for all the TF has been identified, so usually a pseudo-count is used to avoid over fitting of the PWM to the known binding sites only [14, 37, 62]. Intuitively the more the number of known binding sites for a TF, the more reliable is the set of constructed PWMs.

Table 5–6: Comparison between models for the key contributors forming features. aa denotes auto-activation.

Features	Perkins et al.	Jaeger et al.	Trajectory-based	$\beta = 0.1$ Results
A. Hb Peak	Bcd act.	aa+(Bcd +Cad)act.	aa+ (Bcd.+Cad) act.	aa+Bcd act. + Cad rep.
P. Hb Peak	Tll act	Cad Act.	Cad act.	Tll act + Cad Repression
Kr A. Border	hb rep.	(hb+gt)rep.	(hb+gt) rep.	hb rep.
Kr P. Border	Kni rep.	Kni rep.	Kni rep.	(Kni+Gt) rep.
Gt A. Peak	Bcd act.+kr rep.	Bcd act.+kr rep.	Bcd act.+kr rep.	(Bcd +kr) rep.+ Cad act.
Gt P. Peak	Cad act.	Cad act.+tll rep	Bcd +Cad act.,(Kr+ Hb) rep.	Cad act. + aa.
Kni A. Border	Hb + Kr rep.	Hb + Kr rep.	Hb + Kr rep.	Bcd rep.
Kni P. Border	Gt rep.	tll/gt rep.	Gt Rep	tll rep.+ (Cad+Kr) act.

In this section, we devise an optimization strategy that updates the PWMs such that the final cumulative binding site strength data yields a greater level of correlation with the weights obtained from our trajectory-based fits. The main assumption here is that the PWMs used to construct the binding site strength data were not perfect as there is always a possibility of the existence of some unknown binding sites for the TFs. It can be thought of as a method to avoid excessive reliance to the known binding site data while PWM construction.

5.4.1 Error Criteria

We use two error criteria for PWM optimization.

I Proportionality constraint only,

II Proportionality constraint and regularization term.

The first one, the proportionality constraint is the one that we have used in section 5.3. The definition is provided in equation (5.1) and 5.2. The second one adds a regularization term to the definition of E_{prop}. The definition becomes,

$$E_{reg} = E_{prop} + \sum_b c_b^2 \tag{5.3}$$

where,

$$E_{prop} = \sum_b \left(\sum_a \left(\frac{\sum_{a'} |w_{ba'}| \cdot B(b,a')}{\sum_{a'} (B(b,a'))^2} \cdot B(b,a) - |w_{ba}| \right)^2 \right)$$

As c_b is an estimate of the ratio between the w_{ba} and $B(b,a)$, its value increases when the occupancy of binding site decreases. Therefore, the regularization term penalizes any attempt to decrease the strength of the binding site found and awards if the number of binding site hits increases. The rationale behind adding such a regularization parameter is to impose more constraint to a problem which may be under constrained by definition.

5.4.2 Optimization Procedure

We employ a simple randomized local search algorithm for the optimization. The algorithm is described below:

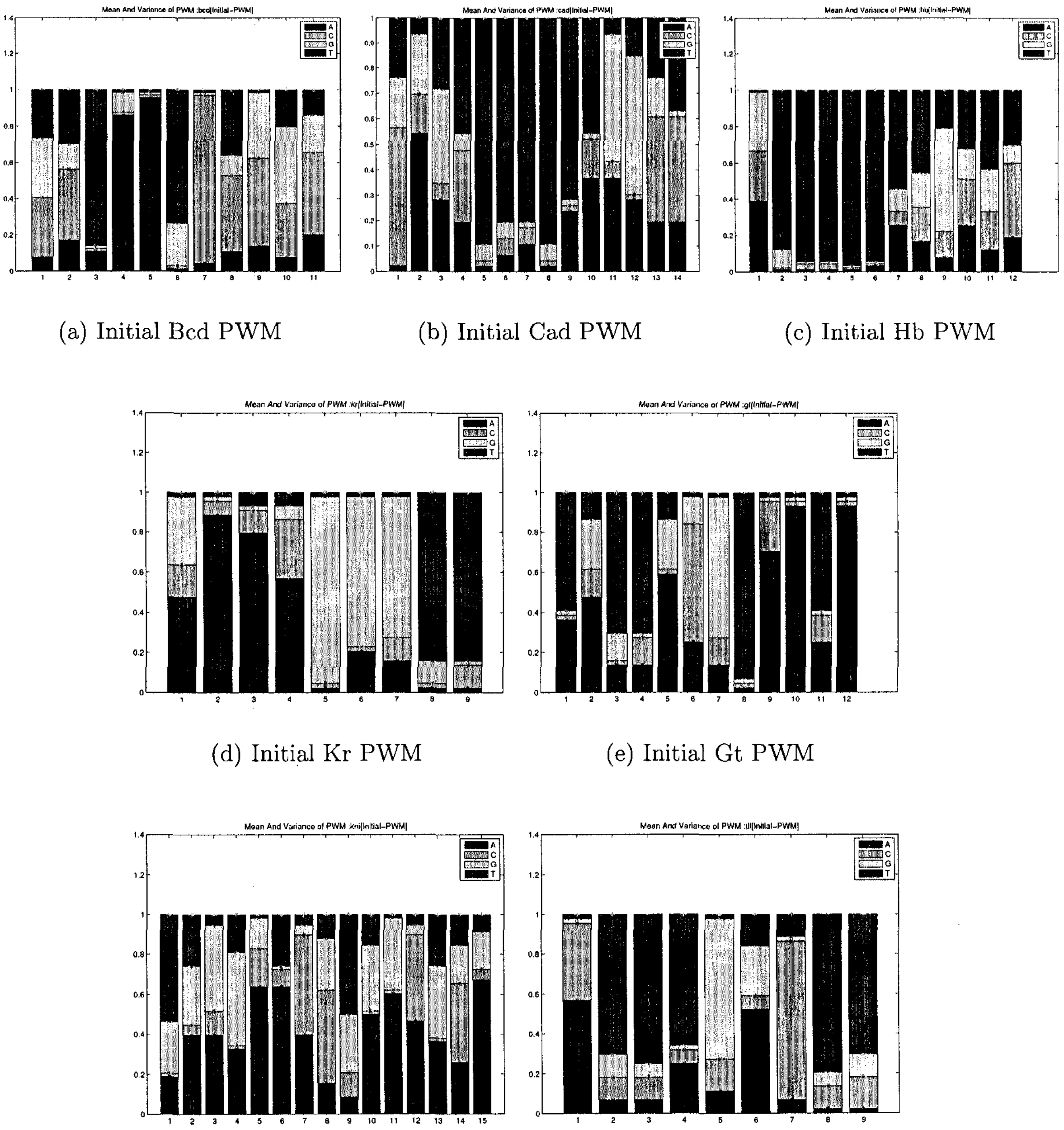
1. Starting from the initial PWMs, perturb a randomly chosen PWM by changing the probability associated with a randomly chosen nucleotide in a randomly chosen position of the PWM.

2. Run Stubb on the changed PWMs to calculate the new CBSS values.
3. Calculate the numeric value of the error criteria. Let's call it E_{new} and the value of the error function at the previous step as E_{prev}.
4. If $E_{new} \geq E_{prev}$ revert the changes to get back to the previous version of the PWM.
5. If $E_{new} < E_{prev}$ then make the changes in PWM permanent.
6. Continue this process until a predefined number of iterations complete.

We have run this algorithm starting from both the Sinha et al. PWMs [58] and a set of random PWMs. The PWMs represents the probability of finding a nucleotide at a given position of the regulatory motif. To perturb a PWM, we choose a random nucleotide at a random position and add or subtract a random number to the corresponding probability value. However, as the sum of the probabilities of all the nucleotides for a given position must always be one, we had to re-normalize the probability values after perturbation. When we start from the given set of PWMs, the amount of change at each step, i.e. the step size parameter, is made inversely proportional to the number of known binding site, as we do not want to make a big change to a PWM which is supported by a large number of known binding sites. The number of iterations were fixed at 10000 for all our optimization runs.

5.4.3 PWM Optimization without regularization

We ran the optimization problem 10 times starting from the Sinha PWMs and 6 times starting from random PWMs. The Sinha PWMs are plotted in Figure 5-8. The results of the optimization runs are decribed below:



(a) Initial Bcd PWM

(b) Initial Cad PWM

(c) Initial Hb PWM

(d) Initial Kr PWM

(e) Initial Gt PWM

(f) Initial Kni PWM

(g) Initial Tll PWM

Figure 5–8: The initial (Sinha) PWMs. Each bar represents the expected frequency of nucleotides at a certain position of the binding site motif. Different nucleotides have been shown with different colours.

Optimization starting from the Sinha PWMs

The error function values are reported in table 5-7. The least error prediction was provided by run no. 8 (5.8 x 10^{-7}) and the largest error prediction was observed for run no. 7 (7.01 x 10^{-5}).

Table 5–7: Error function value after optimization of PWMs starting from the Sinha PWMs

Run No.	Error Function value
1	2.83 x 10^{-6}
2	8.9 x 10^{-7}
3	1.09 x 10^{-6}
4	4.33 x 10^{-6}
5	1.65 x 10^{-6}
6	1.20 x 10^{-5}
7	7.01 x 10^{-5}
8	5.8 x 10^{-7}
9	1.11 x 10^{-6}
10	2.38 x 10^{-5}

Despite the fact that the numeric value of the error function varies for different runs, when we plot the points on a scatter plot as described in subsection 5.3.5, the visual inspection of the plots reveal that for all of the plots, the points approximately lies on a line going through the origin. The scatter plot for Run No. 8 is provided in Figure 5-9 and the scatter plot for Run No. 7 is presented in Figure 5-10. The binding site matrices are recorded in Table 5-8 and 5-9 respectively. The binding site matrices are similar in terms of the magnitudes of the binding site strengths. The one notable exception is that for the first row of the binding site matrix (the weights of Bcd as a TF), the best result PWM(Run 8) yielded almost double hits per target than the worst result PWM (Run 7). The average number of binding sites per target per TF are also approximately equal (Run 7: 2.93, Run 8: 3.10). However, the original CBSS matrix from Sinha et al. detected 3.38 BS per target per TF. So our optimization procedure is in general making the PWMs more restricted in order to reduce the error function.

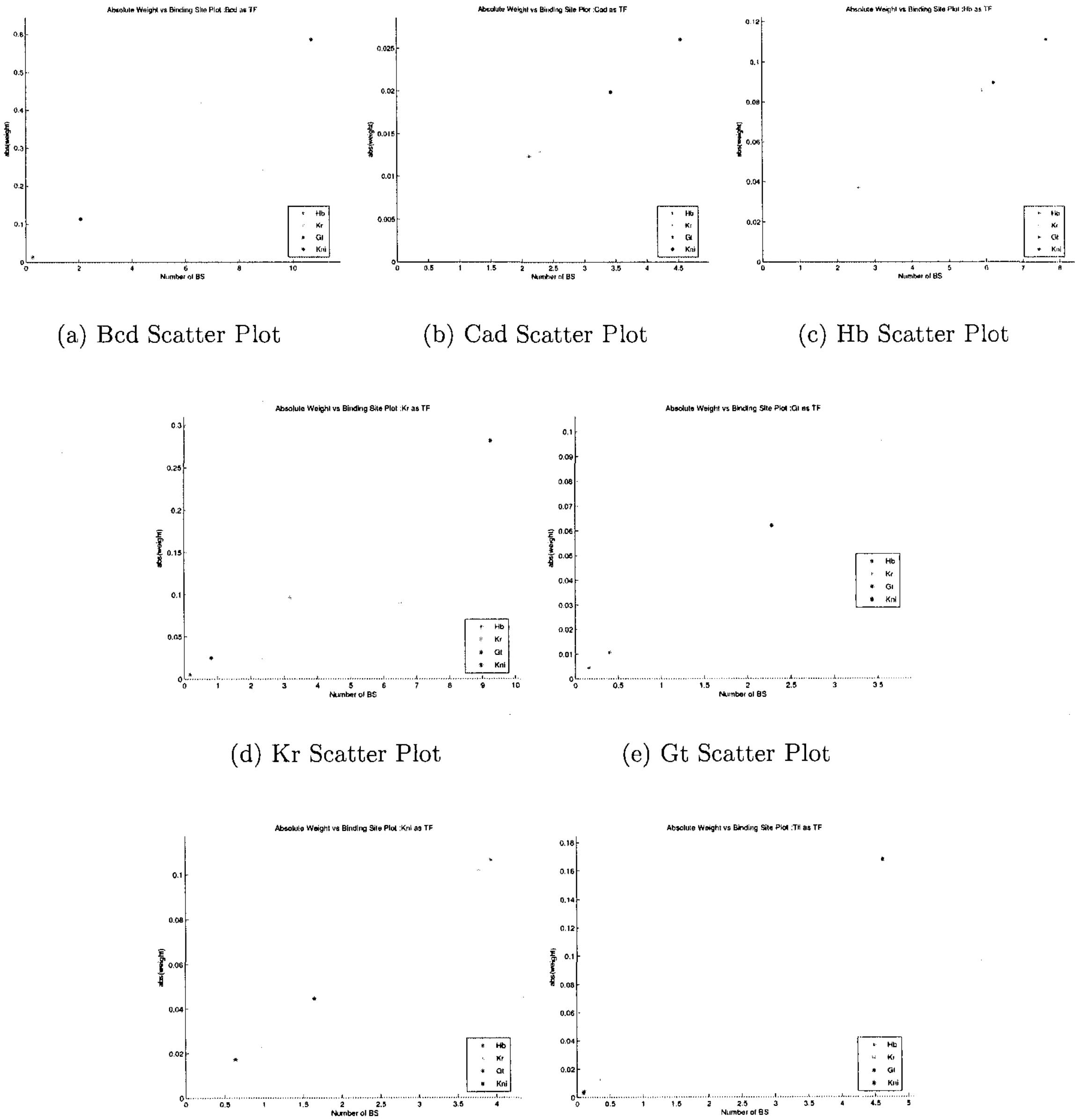
Table 5–8: Cumulative binding site strength obtained from the PWM output of Run 8(The best fit PWM) when optimizing from the given matrices

	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.2567	4.2564	10.6998	2.0771
Cad	2.1090	2.2908	4.5378	3.4258
Hb	2.5468	5.8926	6.2051	7.6545
Kr	0.1755	3.1692	9.2245	0.8082
Gt	0.1609	3.5362	0.3919	2.2782
Kni	3.9349	3.7647	0.6365	1.6509
Tll	0.0902	0.3444	0.1091	4.6127

Table 5–9: Cumulative binding site strength obtained from the PWM output of Run 7(The worst fit PWM) when optimizing from the given matrices

	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.2026	3.2958	8.0011	1.5421
Cad	2.7264	3.1095	6.2348	4.6279
Hb	2.3400	5.4155	5.6529	7.0244
Kr	0.2244	3.4603	10.0813	0.8677
Gt	0.1539	3.2618	0.3641	2.1004
Kni	3.0457	2.9145	0.5227	1.2856
Tll	0.0574	0.2367	0.0861	3.1439

We have also plotted the mean PWMs and the variance observed in Figure 5-11. The figure shows that the variance is the lowest for Hb, and highest for Gt. This is expected as we allow for a greater change in the Gt PWM as it is based on only



(a) Bcd Scatter Plot

(b) Cad Scatter Plot

(c) Hb Scatter Plot

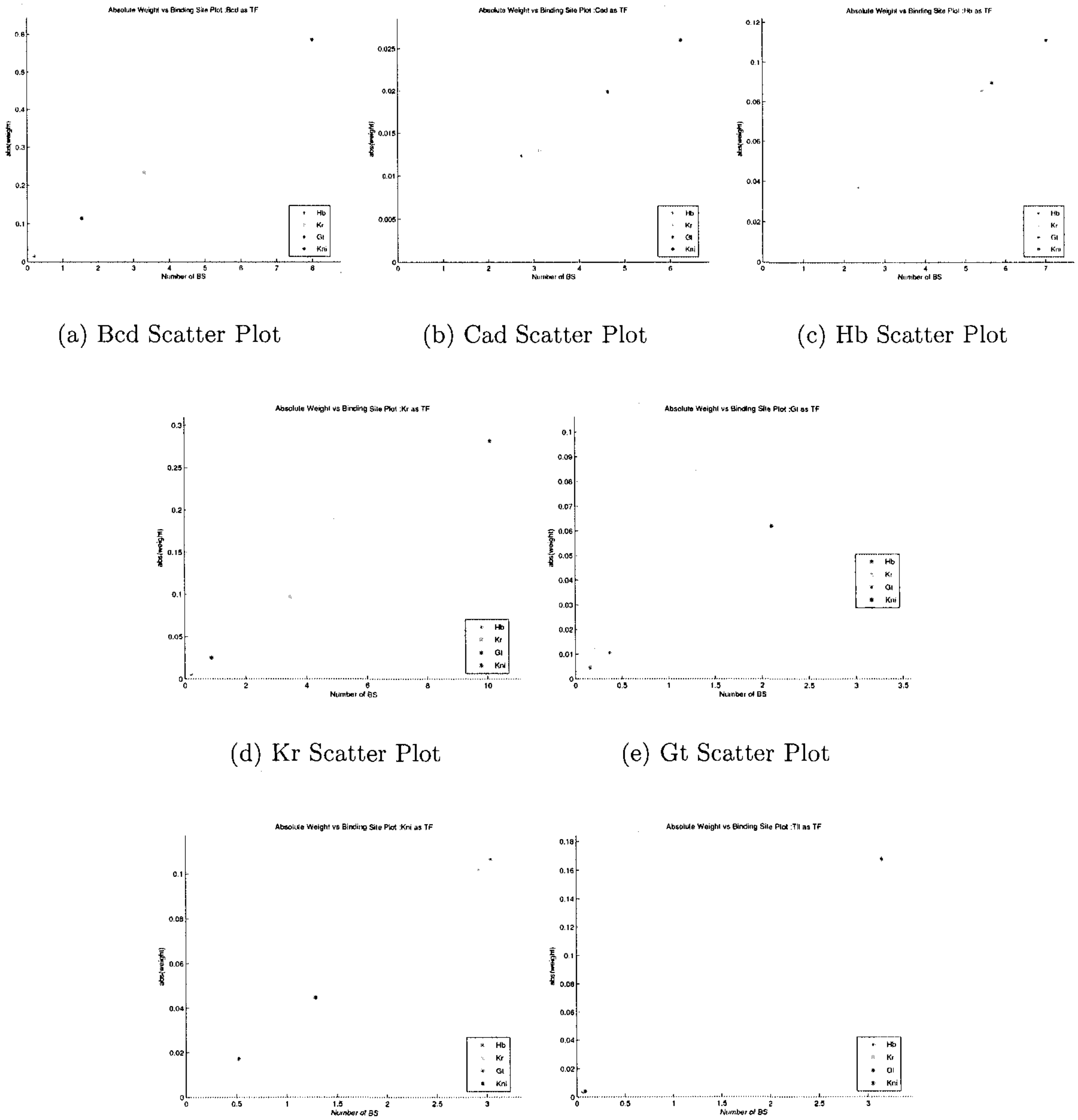
(d) Kr Scatter Plot

(e) Gt Scatter Plot

(f) Kni Scatter Plot

(g) Tll Scatter Plot

Figure 5–9: Binding site strength (X axis) vs regulatory weights (Y axis) for PWM Optimization from the Sinha PWMs(The best fit results)



(a) Bcd Scatter Plot

(b) Cad Scatter Plot

(c) Hb Scatter Plot

(d) Kr Scatter Plot

(e) Gt Scatter Plot

(f) Kni Scatter Plot

(g) Tll Scatter Plot

Figure 5–10: Binding site strength (X axis) vs regulatory weights (Y Axis) for PWM Optimization from the Sinha PWMs(The worst fit results)

8 known binding sites than the Hb PWM, which is supported by 44 known binding sites.

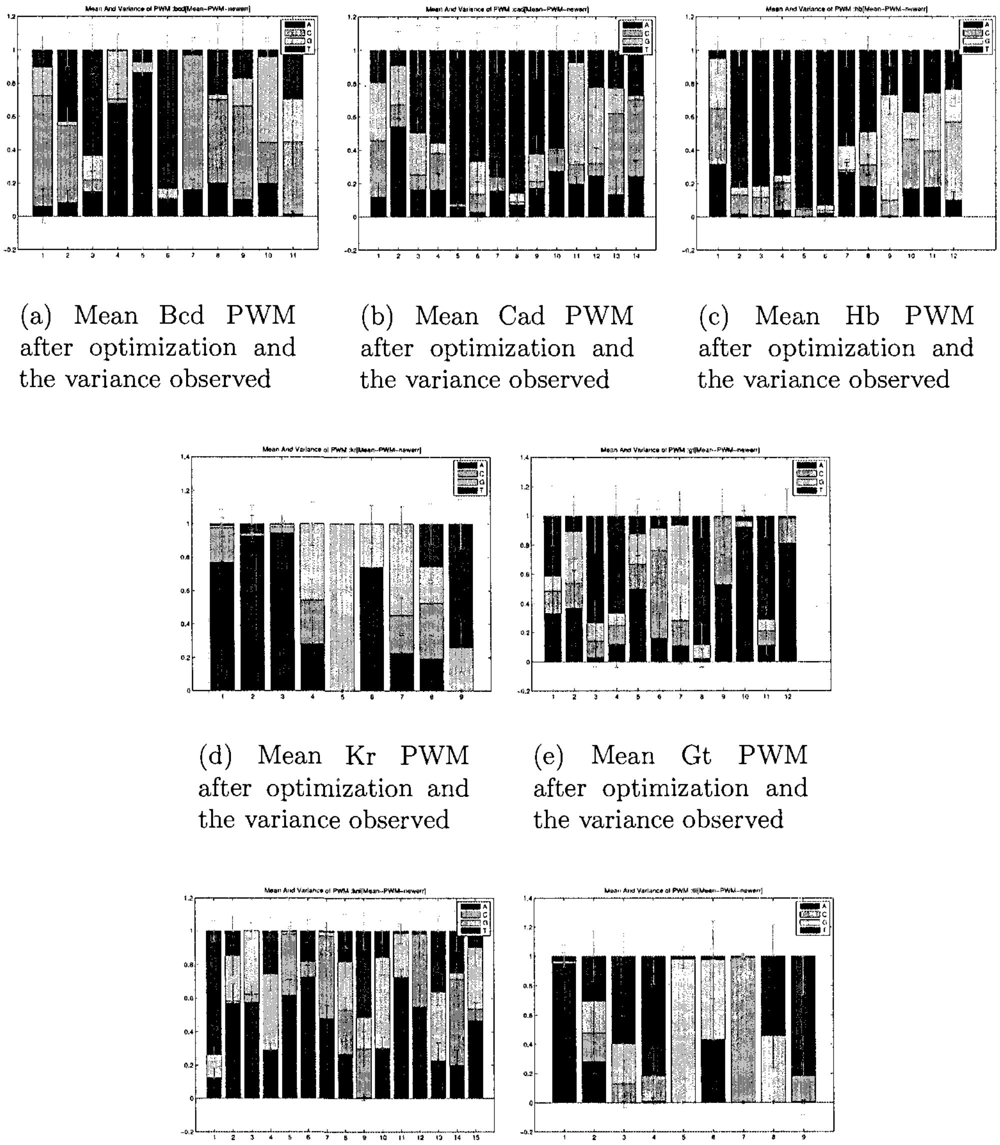
Optimization starting from random PWMs

We initialized the PWMs with random values and let our optimization procedure run for 10000 iterations. The error function values of the optimization results are reported in table TODO. The least error prediction was provided by Run no. 2 (1.01 x 10^{-4}) and the largest error prediction was observed for Run no. 5 (9.4 x 10^{-3}). The error function values have much greater magnitude than the values obtained from PWM optimization from the Sinha PWMs. It infers that when we start from random PWMs, the problem of finding a set of PWM with a greater level of proportionality becomes harder to solve. However, it is not clear how to interpret the numeric values of the error function, we have made scatter plots for run 2 and run 5 to see how the error function value visually affect the relationship between the BS strengths and the regulatory weights.

Table 5–10: Error function value after optimization of PWMs starting from random PWMs

Run No.	Error Function value
1	2.40 x 10^{-3}
2	1.01 x 10^{-4}
3	1.81 x 10^{-3}
4	1.24 x 10^{-3}
5	9.43 x 10^{-3}
6	9.18 x 10^{-4}

The scatter plot for Run No. 5 (the worst fit result) is provided in Figure 5-13 and for Run No. 2 (the best fit results) in Figure 5-12. The binding site



(a) Mean Bcd PWM after optimization and the variance observed

(b) Mean Cad PWM after optimization and the variance observed

(c) Mean Hb PWM after optimization and the variance observed

(d) Mean Kr PWM after optimization and the variance observed

(e) Mean Gt PWM after optimization and the variance observed

(f) Mean Kni PWM after optimization and the variance observed

(g) Mean Tll PWM after optimization and the variance observed

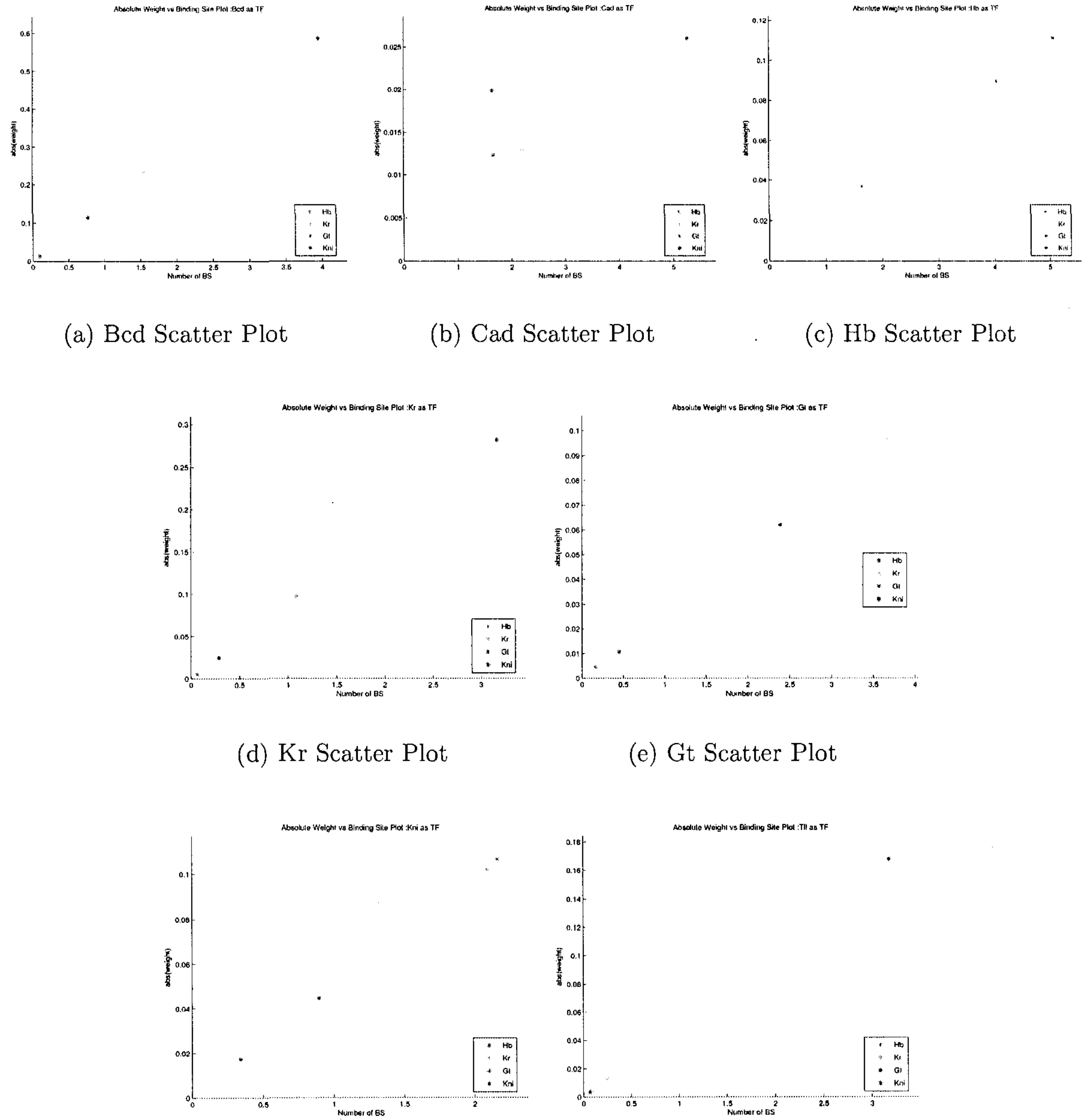
Figure 5–11: Mean PWMs and variance obtained from the optimization runs starting from the Sinha PWMs. Each bar represents the expected frequency of nucleotides at a certain position of the binding site motif. Different nucleotides have been shown with different colours. The variance observed at each position is shown by the error bars.

matrices are recorded in Table 5-12 and 5-11 respectively.The visual inspection of the scatter plots reveal that for all of the plots for Run 2 (the best fit result) except Cad as a TF , the points approximately lie on a line going through the origin. For Cad, we observe that the number of BS detected in the Knirps regulatory region is inadequate to explain the weight of Cad on Kni. However, for Run no. 5, we see that only Hunchback and Knirps nicely satisfy the constraint of proportionality. For the rest of the TFs, there are always some points deviating from the ideal results. The binding site matrices are different from the matrices obtained when we start from the Sinha PWMs. The average number of binding site detected per matrix entry is only between 1.6-1.9 which is much smaller than the optimization results when starting from Sinha PWMs.

Table 5–11: Cumulative binding site strength obtained from the PWM output of Run 2(The best fit PWM) when optimizing from the random matrices

	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.0951	1.5661	3.9479	0.7672
Cad	1.6633	2.2145	5.2626	1.6427
Hb	1.6344	3.8336	4.0439	5.0650
Kr	0.0629	1.0883	3.1607	0.2867
Gt	0.1596	3.6732	0.4454	2.3868
Kni	2.1658	2.0870	0.3416	0.8944
Tll	0.0664	0.2463	0.0745	3.1783

We have plotted the mean PWM and the variance observed in Figure 5-14. The figure shows that the output PWMs are more or less flat (high entropy), i.e. not much variation is observed in the nucleotide frequencies over the positions. As the number of detected binding sites are usually much smaller than the Sinha et al. binding sites, we realize that the algorithm is solving the problem of imposing



(a) Bcd Scatter Plot

(b) Cad Scatter Plot

(c) Hb Scatter Plot

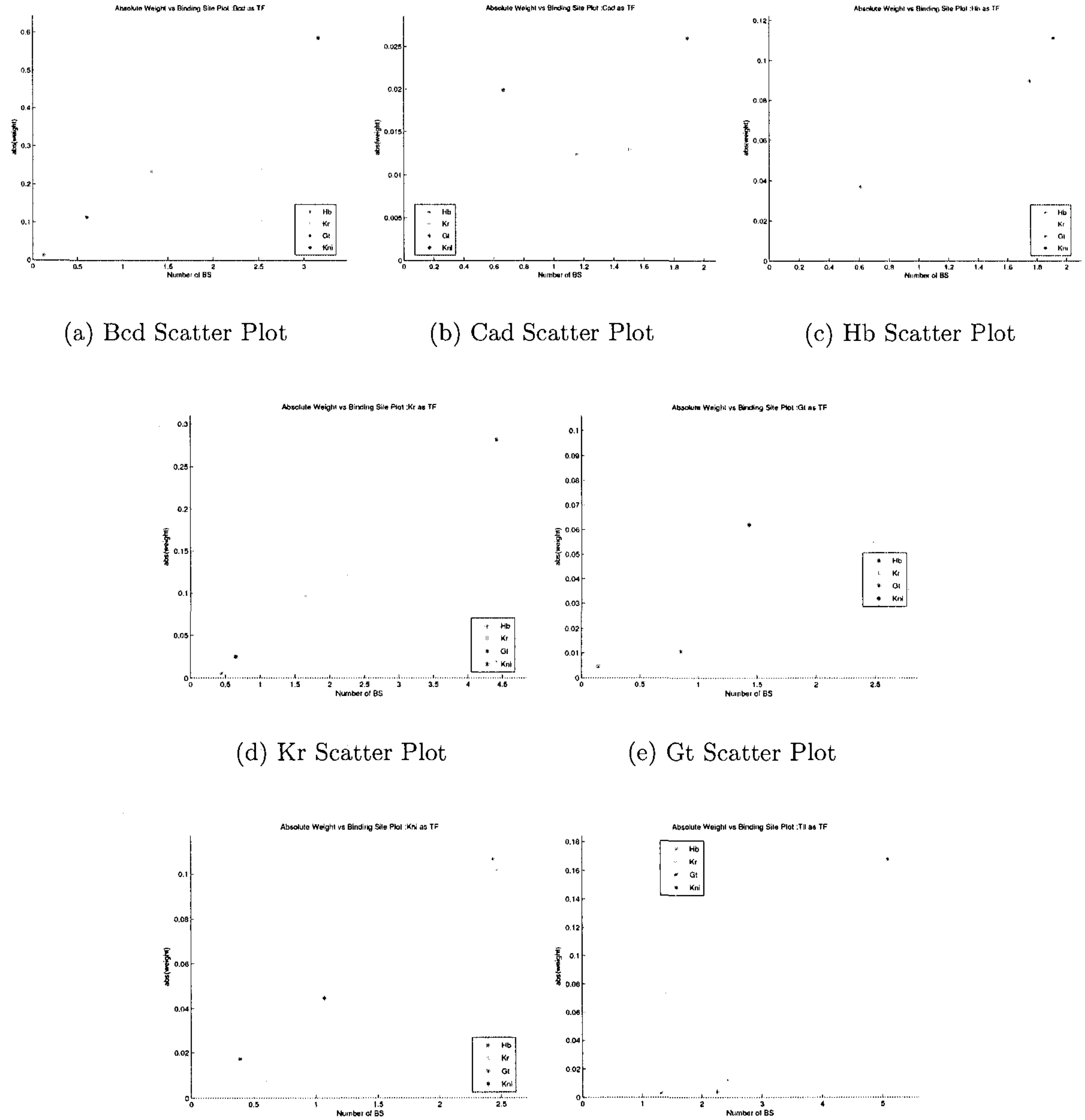
(d) Kr Scatter Plot

(e) Gt Scatter Plot

(f) Kni Scatter Plot

(g) Tll Scatter Plot

Figure 5–12: Binding site strength (X axis) vs regulatory weights (Y axis) for PWM optimization from random PWMs(The best fit results)



(a) Bcd Scatter Plot

(b) Cad Scatter Plot

(c) Hb Scatter Plot

(d) Kr Scatter Plot

(e) Gt Scatter Plot

(f) Kni Scatter Plot

(g) Tll Scatter Plot

Figure 5–13: Binding site strength (X axis) vs regulatory weights for PWM Optimization from random PWMs(The worst fit results)

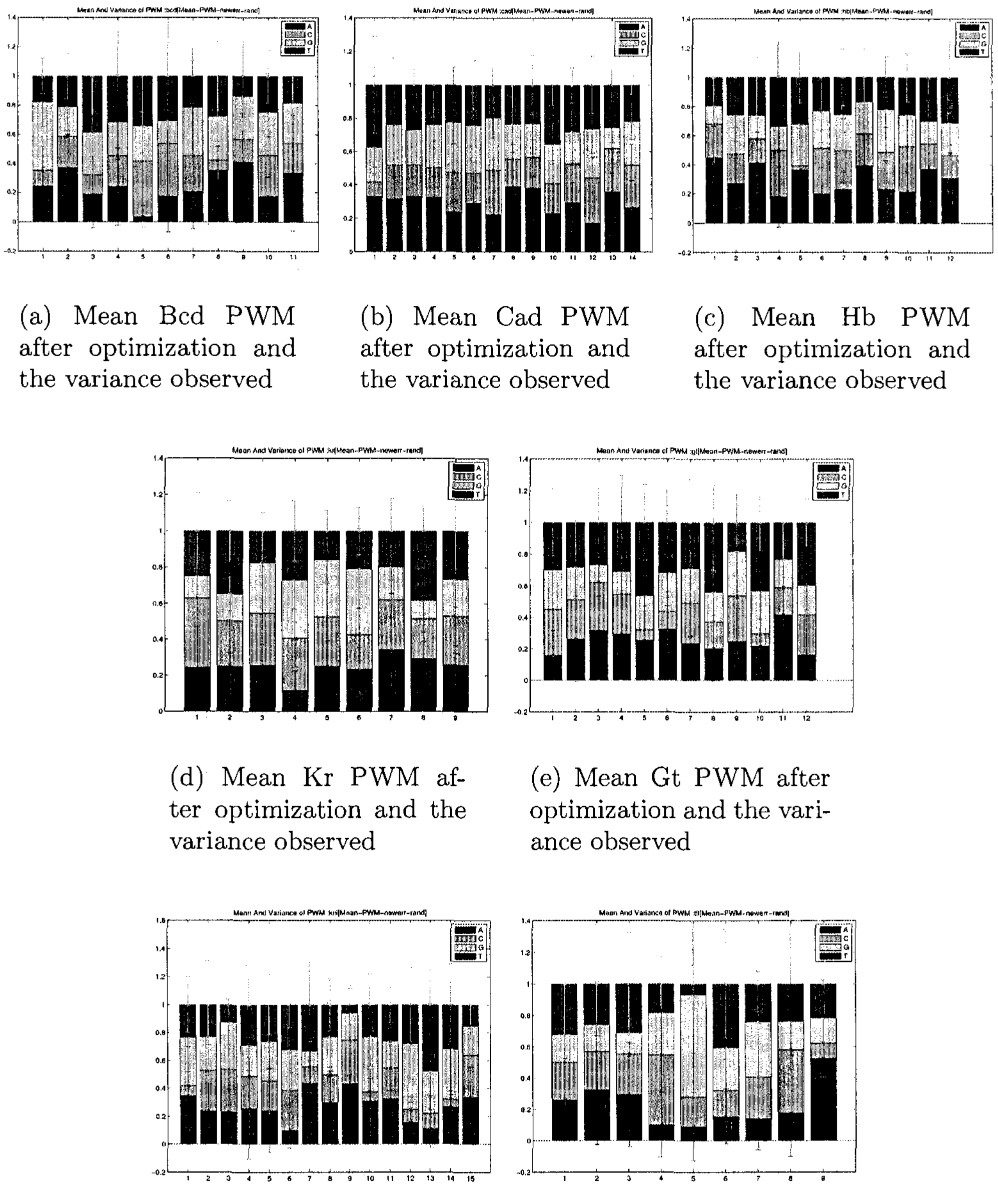
Table 5–12: Cumulative binding site strength obtained from the PWM output of Run 5(The worst fit PWM) when optimizing from the given matrices

	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.1241	1.3295	3.1581	0.6033
Cad	1.1522	1.5003	1.8894	0.6643
Hb	0.6096	1.5990	1.7511	1.9119
Kr	0.4464	1.6410	4.4088	0.6508
Gt	0.1438	2.6014	0.8469	1.4285
Kni	2.4347	2.4659	0.3934	1.0700
Tll	1.3182	2.4261	2.2545	5.0922

the correlation constraint in a rather interesting way. As the algorithm starts with random PWMs which are flat PWMs with not a strong binding affinity to any particular binding sites, when it starts making small changes to the PWMs, it founds out that the proportionality constraint can quite easily be satisfied by just introducing small increase or decrease of the nucleotide frequencies. In that case, the number of binding site hits would not be large, but as our optimization algorithm did not have any constraint on the number of binding site hits, the algorithm does not bother to search beyond this local minima for quest of a new set of PWMs resulting in a CBSS matrix showing a higher level of binding of the PWMs to the regulatory region.

5.4.4 PWM Optimization with regularization

The results from PWM optimization without regularization shows that when we start from the Sinha PWMs, we can find a set of PWMs of the TFs that yield a binding site matrix showing a high level of proportionality with the regulatory weights. However, the average number of binding site hits decreased from the starting point. When we start from random starting points, it is harder to get a good fit and the best of the fits, although visually satisfying the constraints, results in a



(a) Mean Bcd PWM after optimization and the variance observed

(b) Mean Cad PWM after optimization and the variance observed

(c) Mean Hb PWM after optimization and the variance observed

(d) Mean Kr PWM after optimization and the variance observed

(e) Mean Gt PWM after optimization and the variance observed

(f) Mean Kni PWM after optimization and the variance observed

(g) Mean Tll PWM after optimization and the variance observed

Figure 5–14: Mean PWMs obtained from the optimization runs starting from the random PWMs and the variance observed in various positions

flat PWM with a little variation in nucleotide frequency. As a result, the average strength of sites detected decreases to almost half of the same for Sinha PWMs. This phenomena made us believe that the optimization problem itself may be under constrained because we are changing lots of PWM probability parameters to fit a error criteria which is a measure of proportionality between 2 matrices with only 28 elements each. One obvious way to solve the problem is to add new constraints to the optimization criteria. We came up with an idea of adding a regularization parameter emphasizing the need for having stronger binding affinity of the detected binding sites. This may ensure that the average binding strength would increase. The details of this regularization parameter has already been discussed in subsection 5.4.1.

We have run five different optimization runs starting from the Sinha PWMs and five different optimization runs starting from random PWMs with the regularization parameter. The results of the optimization runs are described below:

Optimization starting from the Sinha PWMs

The E_{prop} values are reported in table 5-13. The least error prediction was provided by run no. 1 (4.81 x 10^{-5}) and the largest error prediction was observed for run no. 5 (2.5 x 10^{-4}).

The scatter plots for Run No. 1 is provided in Figure 5-15 and for Run No. 5 in Figure 5-16. The binding site matrices are recorded in table 5-14 and 5-16 respectively. The binding site matrices are very much similar in terms of the magnitudes of the binding site strengths. The average number of binding sites per target per TF are also almost equal. (Run 1: 4.23, Run 5: 4.21). The original CBSS matrix from

Table 5–13: E_{prop} values after optimization with regularization of PWMs starting from the Sinha PWMs

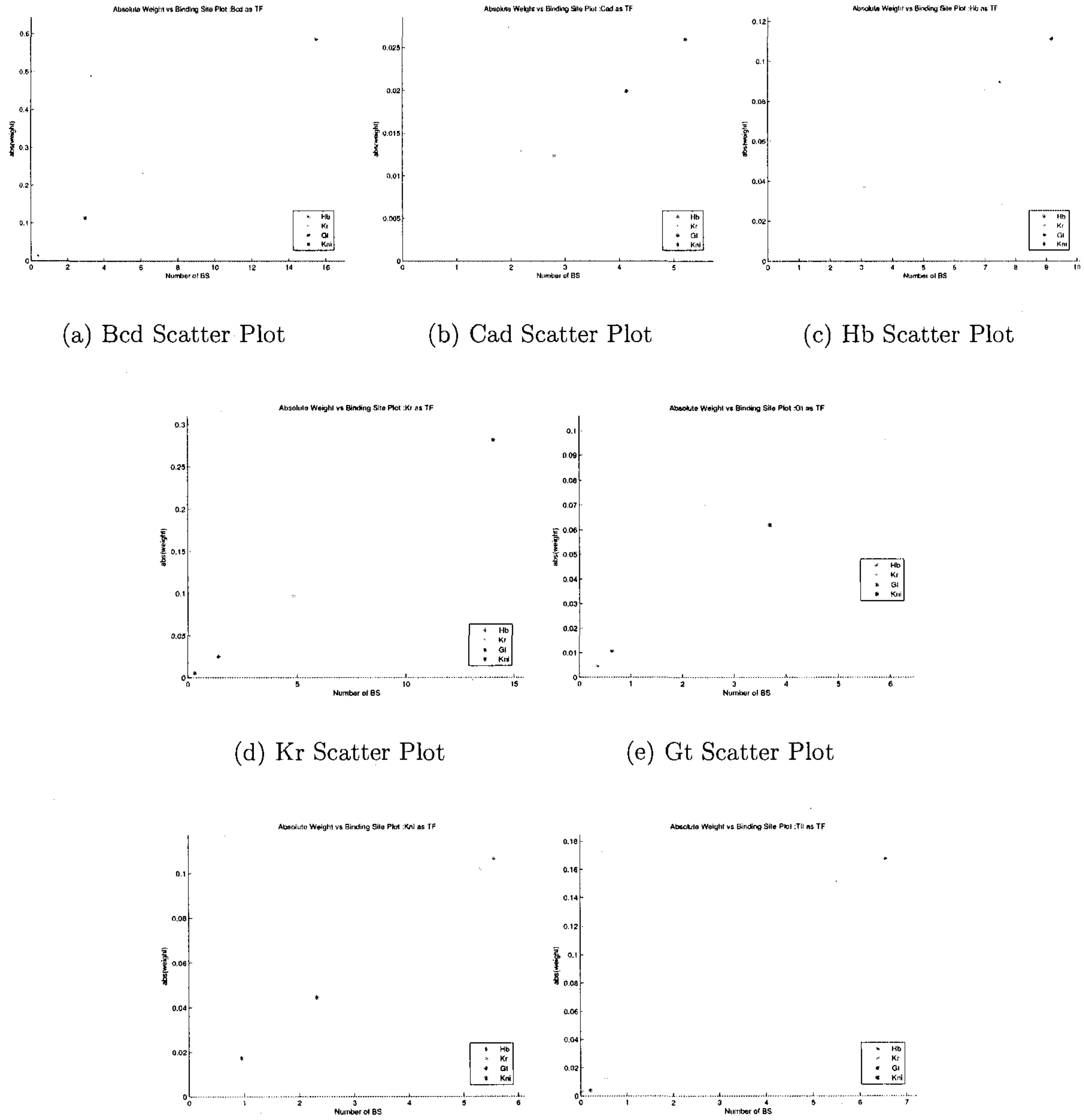
Run No.	**Error Function value**
1	4.81 x 10^{-5}
2	1.36 x 10^{-4}
3	7.72 x 10^{-5}
4	1.173 x 10^{-4}
5	2.5 x 10^{-4}

Sinha et al. detected 3.38 BS per target per TF. So our optimization procedure is by and large increasing the binding site strengths per TF per target due to the new design of the error function.

Table 5–14: Cumulative binding site strength obtained from the PWM output of Run 1(The best fit PWM) when optimizing with regularization from the given matrices

	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.3980	6.0903	15.4943	2.9753
Cad	2.8074	2.1981	5.2175	4.1341
Hb	3.0856	7.0083	7.4854	9.1889
Kr	0.3091	4.8080	14.0419	1.4011
Gt	0.3616	5.8924	0.6367	3.6839
Kni	5.5675	5.3165	0.9513	2.3181
Tll	0.0054	0.5601	0.2111	6.5634

We have also plotted the mean PWM and the variance observed in Figure 5-17. The PWMs underwent a greater degree of change from their initial estimates compared to the optimized PWMs without the use of a regularization parameter. The variance follows the same trend as in the case of simple optimization without regularization, i.e. Hb shows the smallest variance and Gt exhibits the greatest.



(a) Bcd Scatter Plot

(b) Cad Scatter Plot

(c) Hb Scatter Plot

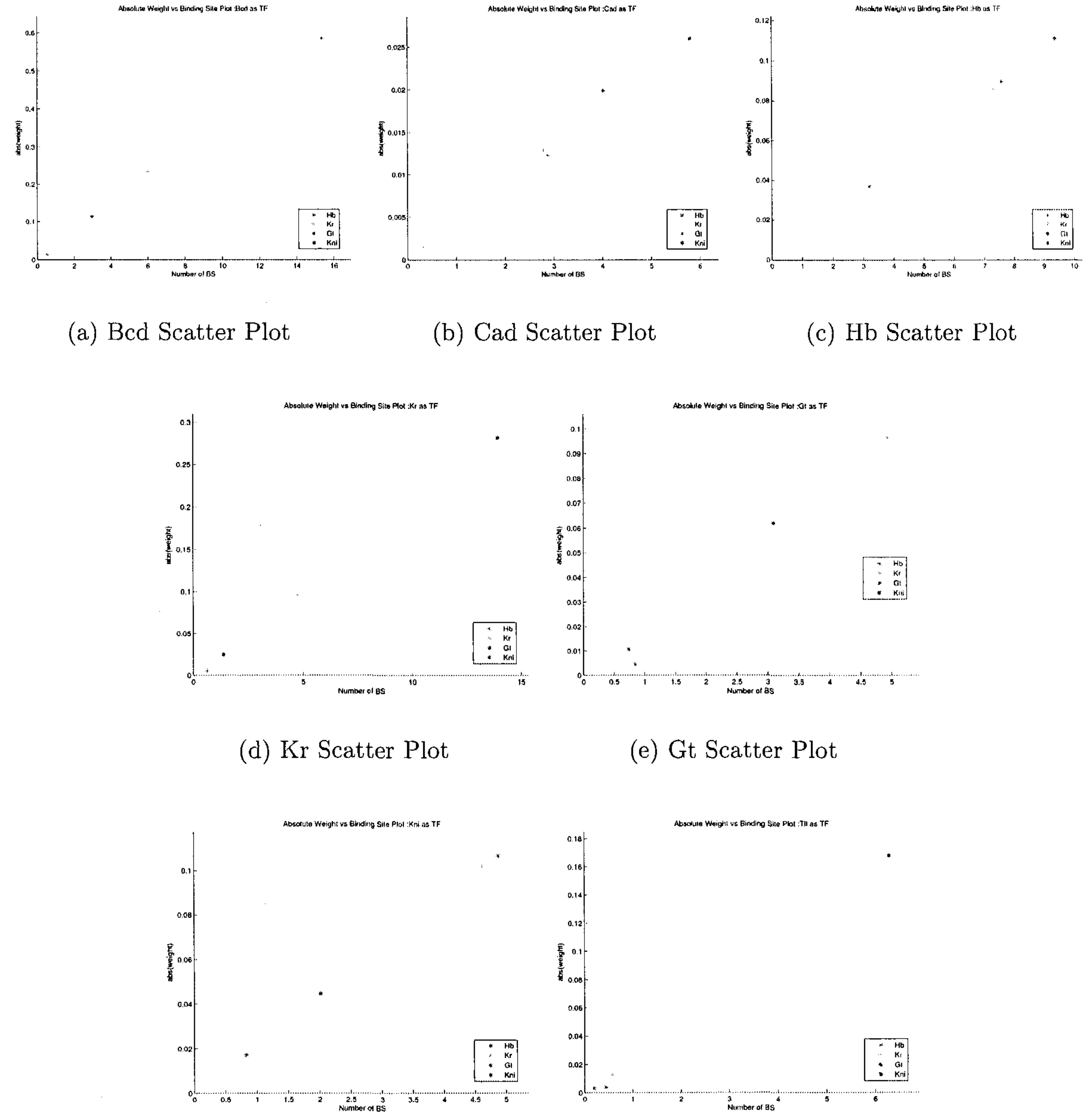
(d) Kr Scatter Plot

(e) Gt Scatter Plot

(f) Kni Scatter Plot

(g) Tll Scatter Plot

Figure 5–15: Binding site strength vs regulatory weights for PWM Optimization from the Sinha PWMs(The best fit results)



(a) Bcd Scatter Plot

(b) Cad Scatter Plot

(c) Hb Scatter Plot

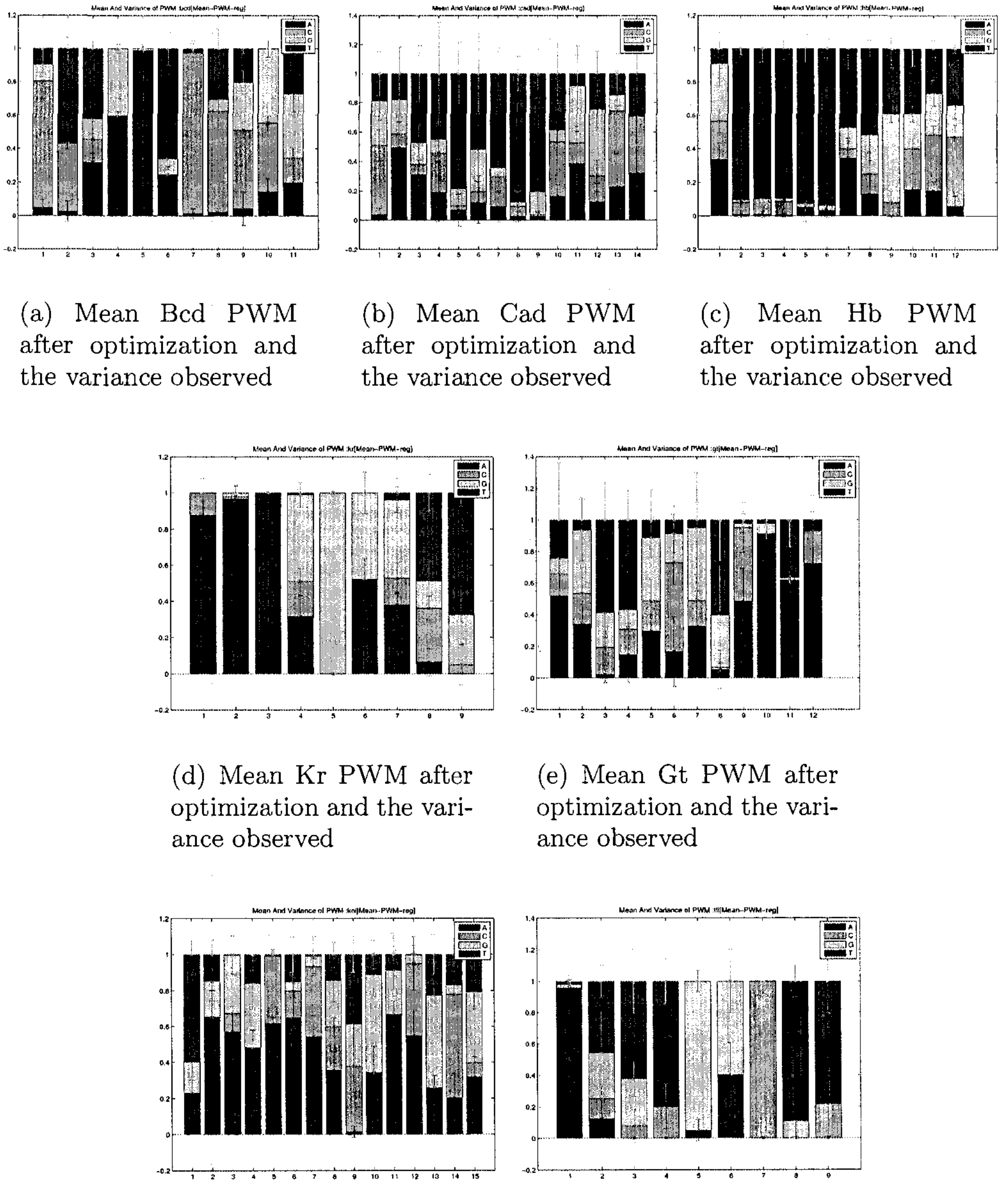
(d) Kr Scatter Plot

(e) Gt Scatter Plot

(f) Kni Scatter Plot

(g) Tll Scatter Plot

Figure 5–16: Binding site strength vs regulatory weights for PWM Optimization from the Sinha PWMs(The worst fit results) with regularization



(a) Mean Bcd PWM after optimization and the variance observed

(b) Mean Cad PWM after optimization and the variance observed

(c) Mean Hb PWM after optimization and the variance observed

(d) Mean Kr PWM after optimization and the variance observed

(e) Mean Gt PWM after optimization and the variance observed

(f) Mean Kni PWM after optimization and the variance observed

(g) Mean Tll PWM after optimization and the variance observed

Figure 5–17: Mean PWMs obtained from the optimization with regularization runs starting from the Sinha PWMs and the variance observed at different positions

Table 5–15: Cumulative binding site strength obtained from the PWM output of Run 5(The worst fit PWM) when optimizing with regularization from the given matrices

	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.5148	6.0172	15.3503	2.9749
Cad	2.8748	2.7895	5.7873	4.0185
Hb	3.1996	7.3024	7.5818	9.3748
Kr	0.6429	4.7261	13.9272	1.3735
Gt	0.8439	4.9416	0.7336	3.0912
Kni	4.8837	4.6174	0.8297	2.0148
Tll	0.1946	0.5614	0.4410	6.2853

Optimization starting from random PWMs

Likewise the case of optimization without regularization, here we also initialized the PWMs with random values and let our optimization procedure run for 10000 iterations. The error function (E_prop) values of the optimization results are reported in Table 5-16. The least error prediction was provided by run no. 4 (1.99 x 10^{-4}) and the largest error prediction was observed for run no. 1 (1.06 x 10^{-3}). The error function values, on average, have greater magnitude than the values obtained from PWM optimization from the Sinha PWMs.We have made scatter plots for Run 1 and run 4 to see how the error function value visually affect the relationship between the BS strengths and the regulatory weights.

Table 5–16: Error function value after optimization with regularization of PWMs starting from random PWMs

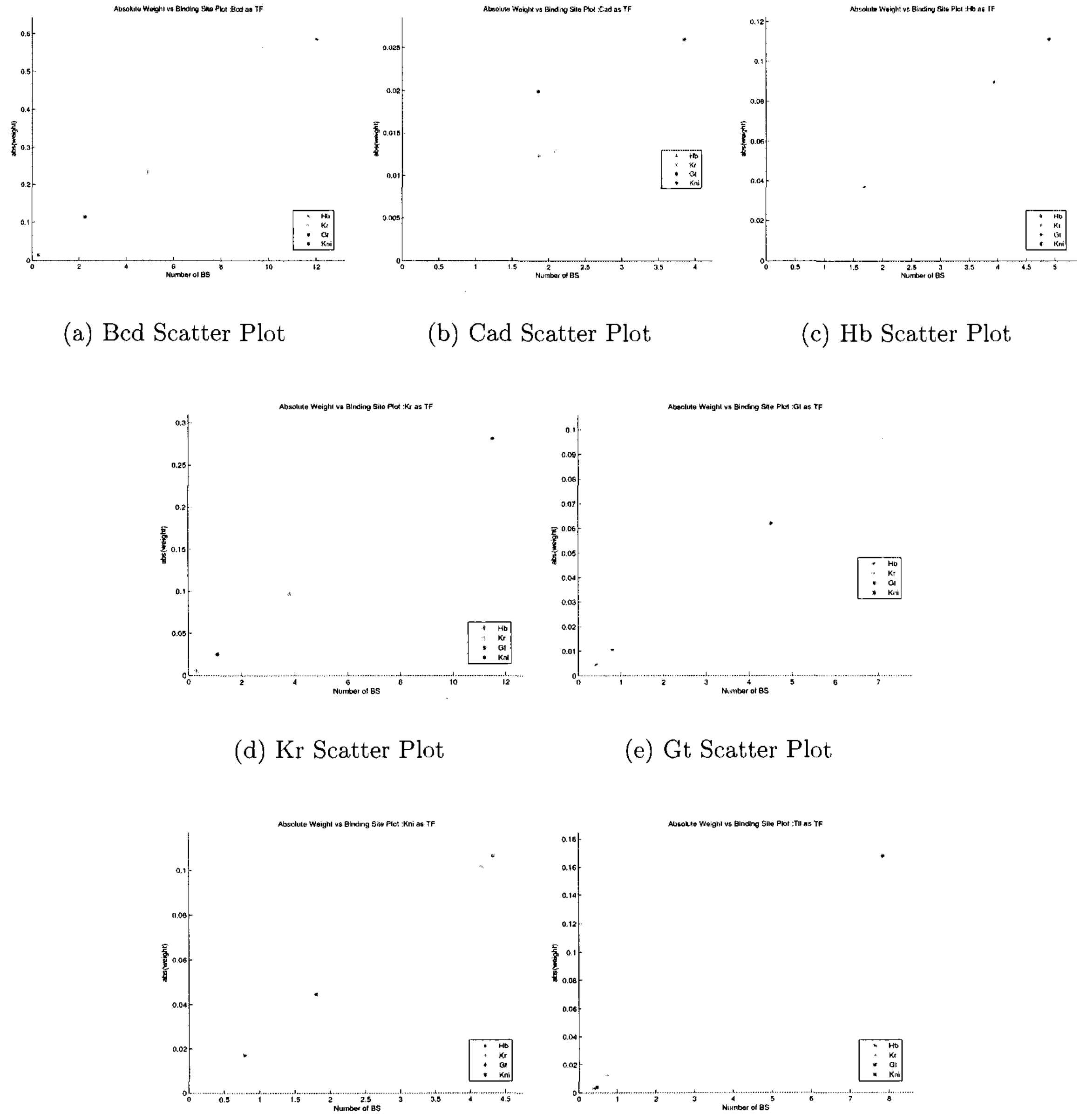
Run No.	**Error Function value**
1	1.0679 x 10^{-3}
2	1.635 x 10^{-4}
3	2.099 x 10^{-4}
4	1.998 x 10^{-4}
5	2.20 x 10^{-4}

The scatter plot for Run No. 1 (the worst fit result) is provided in Figure 5-19 and the same for Run No. 4 (the best fit results) in Figure 5-18. The binding site matrices are recorded in table 5-18 and 5-17 respectively.The visual inspection of the scatter plots reveal that for all of the plots for run 4 (the best fit result) except Cad as a TF , the points approximately lie on a line going through the origin. For Cad, we observe that the number of BS detected in the Knirps regulatory region is inadequate to explain the weight of Cad on Kni. However, for run no. 1 (the worst fit result), we see that only Bcd and Kr nicely satisfy the constraint of proportionality. For the rest of the TFs, there are some points that show deviation from the ideal results. The binding site matrices are different. The average number of binding site detected per TF per target is only 2.89 for run 1. But for run 4, the average goes up to 3.40.

Table 5–17: Cumulative binding site strength obtained from the PWM output of Run 2(The best fit PWM) when optimizing from the random matrices

	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.2651	4.8935	12.0124	2.2684
Cad	1.8595	2.0818	3.8495	1.8584
Hb	1.6827	3.7733	3.9388	4.8969
Kr	0.3049	3.8089	11.5105	1.0896
Gt	0.4087	7.1010	0.7963	4.5120
Kni	4.3259	4.1678	0.7874	1.7970
Tll	0.3918	0.7423	0.4790	7.8471

We have also plotted the mean PWM and the variance observed in Figure 5-20. The flatness of the PWMs decrease considerably when we compare them to the results of the corresponding case without regularization. Still, the entropy is higher than the results obtained when we start from the Sinha PWMs as opposed to random



(a) Bcd Scatter Plot

(b) Cad Scatter Plot

(c) Hb Scatter Plot

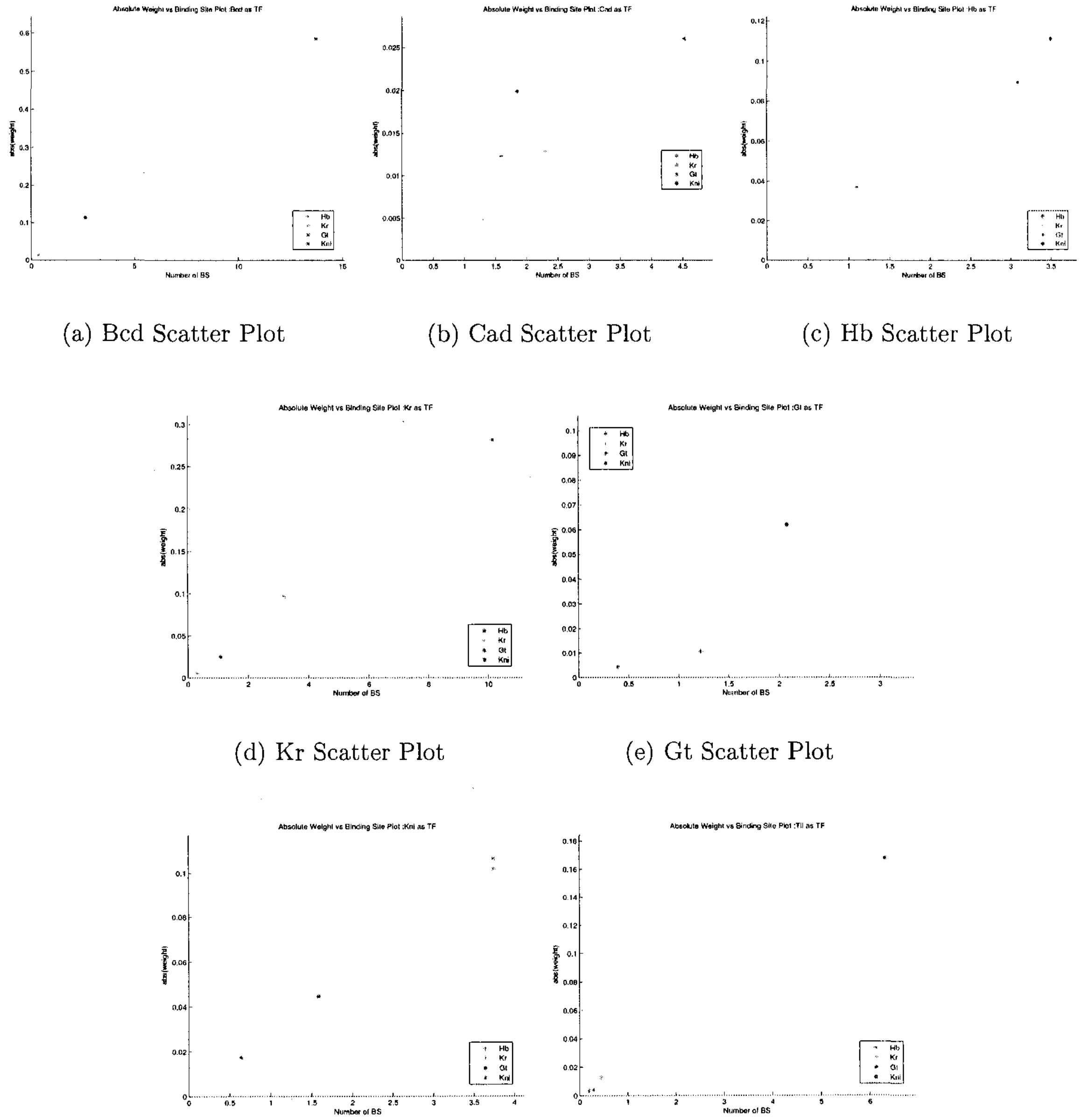
(d) Kr Scatter Plot

(e) Gt Scatter Plot

(f) Kni Scatter Plot

(g) Tll Scatter Plot

Figure 5–18: Binding site strength (X axis) vs regulatory weights (Y axis) for PWM optimization with regularization from random PWMs(The best fit results)



(a) Bcd Scatter Plot

(b) Cad Scatter Plot

(c) Hb Scatter Plot

(d) Kr Scatter Plot

(e) Gt Scatter Plot

(f) Kni Scatter Plot

(g) Tll Scatter Plot

Figure 5–19: Binding site strength (X axis) vs regulatory weights (Y axis) for PWM optimization with regularization from random PWMs(The worst fit results)

Table 5–18: Cumulative binding site strength obtained from the PWM output of Run 5(The worst fit PWM) when optimizing from the given matrices

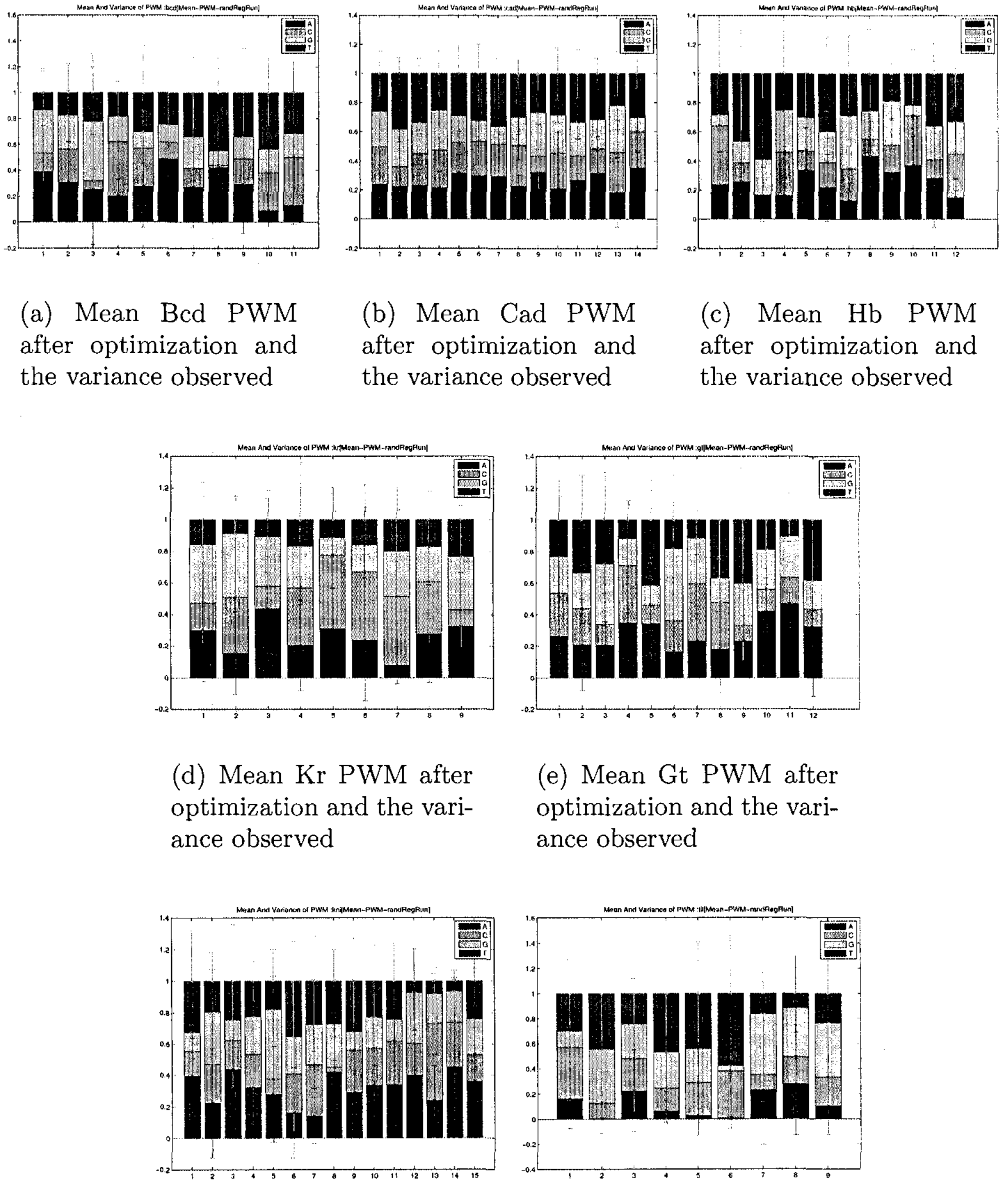
	Hb	**Kr**	**Gt**	**Kni**
Bcd	0.3506	5.4789	13.7384	2.6432
Cad	1.5987	2.3117	4.5341	1.8540
Hb	1.0992	2.5011	3.0870	3.4943
Kr	0.2973	3.1754	10.1516	1.0872
Gt	0.3956	3.0358	1.2188	2.0795
Kni	3.7412	3.7463	0.6424	1.5856
Tll	0.1963	0.4456	0.2953	6.3023

PWMs. The variance of the frequencies is also higher than the other experiments. The average number of binding site detected is also greater than the Sinha BS. But the number of binding site is less compared to the optimization results obtained when we start from the Sinha PWMs.

5.4.5 Distance between different PWMs

A PWM collection is the set of PWMs (each corresponding a particular TF) obtained from a particular experiment setting. There are several possible way to calculate the distance between two PWM collections. We have used a very simple distance measure that takes into account shifted alignments and the reverse complement matches. Let's assume P and Q are two different PWM collections, both of which include N PWMs $P_{1 \ldots N}$ and $Q_{1 \ldots N}$, one for each TFs. Then the distance between P and Q is calculated using the following equation:

$$Dist(P,Q) = \frac{\sum_i min\left(Dist(P_i, Q_i), Dist(P_i, \bar{Q}_i)\right)}{N} \tag{5.4}$$



(a) Mean Bcd PWM after optimization and the variance observed

(b) Mean Cad PWM after optimization and the variance observed

(c) Mean Hb PWM after optimization and the variance observed

(d) Mean Kr PWM after optimization and the variance observed

(e) Mean Gt PWM after optimization and the variance observed

(f) Mean Kni PWM after optimization and the variance observed

(g) Mean Tll PWM after optimization and the variance observed

Figure 5–20: Mean PWMs obtained from the optimization with regularization runs starting from the random PWMs and the variance observed in various positions

and,

$$Dist(P_i, Q_i) = min_{-3 \leq l \leq 3} \sum_{j=1}^{length(P_i)} \frac{\sum_{k=1}^{4} (P_{ijk} - Q_{i(j-l)k})^2}{length(P_i)} \tag{5.5}$$

where, P_{ijk} and Q_{ijk} are the probability of observing nucleotide k at the position j of the regulatory motif belonging to TF i. Equation (5.5) represents the distance measure between two Sinha PWMs. The distance is just the sum of squared difference between the probability values in the two matrices at the corresponding positions. We consider a shift of up to three nucleotides at both side (the l variable in Equation (5.5)) and take the minimum of all these distances. This is to ensure more flexibility while comparing different PWMs. The distance between matrix P_i and the reverse complement of $\bar{Q}_i$, constructed by reversing the Q_i and replacing nucleotides with their reverse complements (A with T, C with G and vice versa), is also measured. The final distance between P and Q is just the mean distance of a matrix in P with the corresponding matrix in Q or its reverse complement.

This distance measure gives equal weight to all the TFs and it does not depend on the motif length (as we take the average by dividing the squared error by its length). Moreover, consideration of shifts and reverse complement augments the match criteria to make it more realistic measure of dissimilarity. Figure 5-21 includes an image plot of the mean distance between different PWM collections.

Our analysis of the plot reveals:

I The distance among the optimization runs without regularization starting from the Sinha PWMs are the smallest compared to other distances. These distances are less than the distances between the initial PWM and the optimized PWMs.

We can deduce that all the PWM collections after optimization without regularization from the Sinha PWMs are very similar and the optimization runs are essentially converging to a neighborhood of the search space which is not too far away from their starting point.

II When we add the regularization parameter, the solution obtained are different from the solution obtained without regularization. However, the distance from the initial PWMs are of the same order. Due to the introduction of the regularization parameter, now the optimization runs try to get as much hits as possible. The distance results infer that the initial PWMs does not need to be changed too much to find such solutions. These solutions cover a slightly greater region in the search space.

III The random PWM results are far away from the Sinha PWMs and the optimization results from Sinha PWMs. Adding regularization parameters push them further away from each other as well as from the Sinha PWMs. The optimization results exhibit more inter-distance than the case of starting from good PWMs. It may imply that the problem may still be under constrained, i.e. there are many optimal or near optimal solutions which may be reached easily from most random starting points.

IV The distance between the optimization results starting from random PWMs increases over the iterations. It shows that the optimization procedure, when starting from random PWMs, can find some reasonable solutions by taking almost any path from the random starting points.

5.5 Summary of our Findings

We summarize the results of associating the binding site data with the regulatory weights below:

i Our attempt to optimize the weights to find a good proportionality component with the binding site data fails to fit the data well. The regulatory interactions are also not identified properly.

ii The reverse optimization problem of finding a new set of PWMs to generate a new cumulative binding site strength values showing a greater level of proportionality with the binding site data can be solved with a reasonable level of accuracy. However, if no regularization parameter is used, the general tendency of such an optimization procedure is to cut down the binding strengths which is apparently making it easier to solve the problem.

iii The introduction of the regularization parameter in the error criteria of the optimization procedure naturally increases the overall binding of the factors to the promoter.

iv When we start from random PWMs instead of the Sinha PWMs, the optimization results in flat PWMs. Adding a regularization parameter contributes towards decreasing the flatness of the PWMs. Nevertheless, the PWMs still show greater level of entropy compared to the results of PWM optimization from the Sinha PWMs.

v The analysis of the inter-distance of the optimization results reveals that the problem formation, even with regularization, may still be under-constrained.

vi Despite the under-constrained nature of the problem, the PWM optimization from the Sinha PWMs perform better than the random PWM optimization in terms of minimization of the error function. It may be an indication that the initial PWMs are by and large correct, and making small changes to them can lead us to a reasonable solution. This is a positive sign, because the known PWMs are derived from known binding sites of a factor and it is highly unlikely that the true PWM would be completely different from the known one.

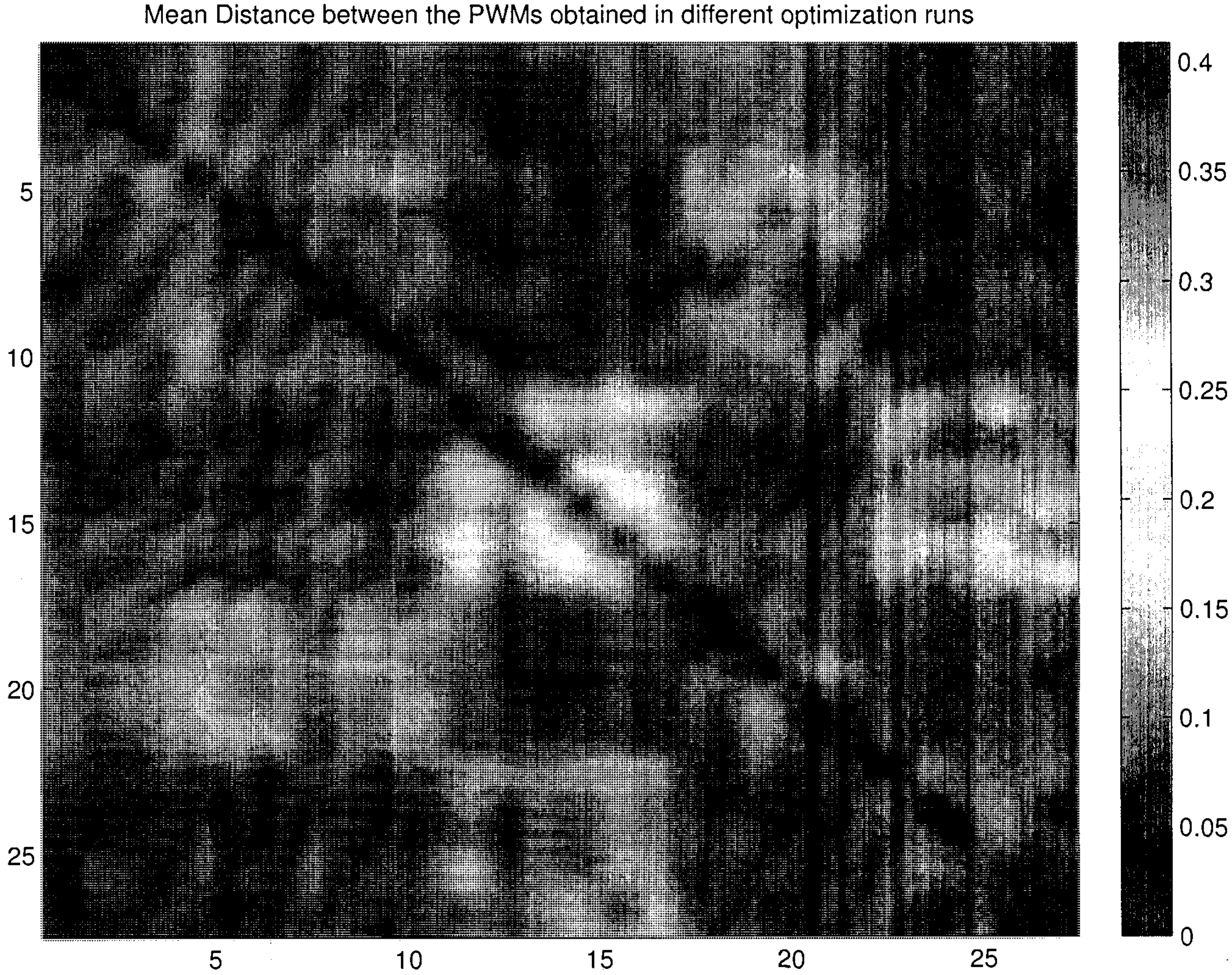


Figure 5–21: The Distance between different PWM Collections. 1-10 are the ten optimization without regularization results starting from the Sinha PWMs, 11 is the Sinha PWMs, 12-17 are six optimization without regularization results starting from random PWMs, 18-22 are five optimization with regularization results starting from the Sinha PWMs and 23-27 are five optimization with regularization results starting from random PWMs

CHAPTER 6
Discussion and Conclusions

In this thesis, we focus our studies on two related problems. The first problem is to model the gene regulatory network of the *Drosophila melanogaster* to infer the regulatory relationship governing the expression profile of the gap genes. This problem is widely addressed in the literature. We had started with a very simple static model in Chapter 3 which, despite its simplicity, had been able to correctly predict the modes (activation/ repression) of most of the important interactions. Then, in Chapter 4, we extended our model to include the time series expression profile which results in the dynamical models. Instead of the usual approach of modelling the dynamical data using differential equation model, we devised a discrete-time gene circuit model that was successful in reconstructing the expression profile with a greater degree of accuracy and was also able to capture the prominent interactions in the network. We tested two slightly different discrete time models, namely, transition-based models and trajectory-based models. Transition-based models,which could explain the formation of many important features in the expression of the genes, are the direct extension of the static models . The trajectory-based models are more biologically accurate models. These models also performed the best on the given data in terms of the reduction of error. All the models had difficulties to predict the expression profiles at the earlier time steps, but this problem is a common feature for all the gene circuit models in literature. In fact, a recent study [47] has shown that the early activation of the gap genes cannot be fully determined by the known morphogens like Bcd and Hb alone. It proposes that some additional regulating factors affect the early *Drosophila* expression.

The second problem that we address here is the fundamental problem of relating the expression profile with the inherent cause of gene regulation, namely, the binding of the transcription factors to their respective binding sites which are located in the target gene regulatory region. The cumulative binding site composition (CBSS) can be determined if the position weight matrices (PWM) for all the transcription factors are known correctly. The regulatory weights can be determined from the expression data using various techniques and modelling. Unfortunately neither the binding site data nor the regulatory weights data is fully reliable, because there is no universally agreed upon model to describe and determine the *true* values for these two factors. Therefore, in Chapter 5, we had designed two optimization problems each of which optimizes one dependent factor (either regulatory weights or PWM/BS data) in accordance with an assumed relationship of proportionality with the other factor which is held fixed. We define the notion of proportionality which extends the idea of standard definition of correlation to impose a strict proportionality constraint between these two factors. However, when we try to optimize the weights by keeping the binding site data fixed, we get a regulatory weight matrix incapable of explaining the true causes of gene expression profile despite its success in reproducing all the principal domains of the expression profile. On the other hand, when we optimize the PWM data to find a binding site composition retaining a high level of proportionality, the optimization problem can be solved up to a reasonable level of accuracy. However, further analysis on the result demonstrated that this optimization problem may be under-constrained and so, the optimized PWMs may be subject to overfitting problem. Using a regularization parameter imposed more restrictions on the problem definition, yet analysis revealed that although it made some progress towards solving the problem, it alone was not sufficient to get rid of the under-constrained nature of the problem.

In our opinion, the future works on this subject should be directed towards combatting the under-constrained nature of the PWM optimization problem. There are several possible paths of doing so. The most obvious one is to change the way we calculated the binding site strengths from the Sinha PWMs and promoters. The Stubb algorithm used by Sinha et al. [59] considers the competitive binding that exists between different transcription factors, but it does not model the quenching effect [21, 20] , silencing effect [4], cooperativity and the thermodynamic kinetics of a ligand bound to a binding site. A more realistic model would take these factors into consideration while computing the CBSS values from the given sequence information. A recent study [31] proposes a data driven approach which considers these factors to predict the expression a stripe of even skipped (*eve*) gene from the sequence and the quantitative gene expression data. It also shows that a CRM driving a specific pattern of expression is not necessarily a continuous sequence of nucleotides containing a compact arrangement of clustered binding site, it can be a diffuse CRM whose binding sites are distributed over a large DNA segment. To extend this model to cover all the gap genes simultaneously for all the time points would be an interesting task. Reconstructing the gap gene expression data directly from the gene sequence information and initial expression profile of the morphogens using a single optimization procedure is certainly a more challenging problem. We believe a probabilistic model that includes details of the regulatory mechanism at a physiological level may be more appropriate for incorporating into our works.

Another potential direction the future researcher may explore is imposing a more realistic assumption on the precise relationship between the binding site strengths and the regulatory weights. In this work, we have proposed a very simple linear model to represent this relationship. A soft threshold based model may be more biologically authentic. The missing morphogens determining the early time step expression profile, if identified, may improve the quality of prediction further and may result in a better regulatory weight matrix.

The problem of optimizing the weights with the binding site data set fixed was based on the assumption of fixed TF potency. The other natural direction would be to assume that each target has a fixed *sensitivity* parameter. The most realistic model should consider both the TF potency and target sensitivity assumption. Further research can be conducted on this problem as well.

Finally, we have used a very simple randomized local search technique for the PWM optimization problem. While it looks sufficient for our formulation of the problem, it might not be good enough for a newly designed problem imposing more constraints on the optimization criteria. Therefore, researchers must also devise a better optimization technique for solving these problems. While it is not possible to cross-validate the regulatory weight regression problem due to the lack of i.i.d samples, it is possible to cross-validate the PWM optimization procedure by constructing PWMs based on a subset of the known binding sites and then evaluating the capability of the optimized PWM to predict the left out binding sites. Cross validating the results can give the researcher a fair idea of the performance of the optimization runs.

References

[1] M. Akam. The molecular basis for metameric pattern in the Drosophila embryo. *Development*, 101:1–22, 1987.

[2] A. Arkin, J. Ross, and H.H. McAdams. Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage λ-Infected Escherichia coli Cells. *Genetics*, 149(4):1633–1648, 1998.

[3] T.L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21(1):51–80, 1995.

[4] S. Barolo and M. Levine. hairy mediates dominant repression in the Drosophila embryo. *The EMBO Journal*, 16:2883–2891, 1997.

[5] L. Bintu, NE Buchler, HG Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005.

[6] L. Bintu, NE Buchler, HG Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev*, 15(2):116–24, 2005.

[7] M. Blanchette, A.R. Bataille, X. Chen, C. Poitras, J. Laganière, C. Lefèbvre, G. Deblois, V. Giguère, V. Ferretti, D. Bergeron, et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Research*, 16(5):656–668, 2006.

[8] H. Bolouri and E.H. Davidson. Modeling transcriptional regulatory networks. *BioEssays*, 24(12):1118–1129, 2002.

[9] H.J. Bussemaker, H. Li, and E.D. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27:167–174, 2001.

[10] M. Capovilla, E.D. Eldon, and V. Pirrotta. The giant gene of Drosophilaencodes a b-ZIP DNA-binding protein that regulates the expression of other segmentation gap genes. *Development*, 114:99–112, 1992.

[11] K.W. Chu, Y. Deng, and J. Reinitz. Parallel simulated annealing by mixing of states. *J Comput Phys*, 148:646–662, 1999.

[12] DE Clyde, MS Corado, X. Wu, A. Pare, D. Papatsenko, and S. Small. A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. *Nature*, 426(6968):849–53, 2003.

[13] E.H. Davidson. *Genomic Regulatory Systems: Development and Evolution.* Academic Press San Diego, CA, 2001.

[14] IB Dodd and JB Egan. Systematic method for the detection of potential lambda Cro-like DNA-binding regions in proteins. *J Mol Biol*, 194(3):557–64, 1987.

[15] R. Durbin, S. Eddy, A.S. Krogh, and G. Mitchison. Biological sequence analysis: Probabilistic models of proteins and nucleic acids, 1998.

[16] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns, 1998.

[17] E.D. Eldon and V. Pirrotta. Interactions of the Drosophila gap gene giant with maternal and zygotic pattern-forming genes. *Development*, 111:367–378, 1991.

[18] ZH Gerald and DS Gary. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences [J]. *Bioinformatics*, 15(7/8):563–577, 1999.

[19] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[20] S. Gray, H. Cai, S. Barolo, and M. Levine. Transcriptional Repression in the Drosophila Embryo. *Philosophical Transactions: Biological Sciences*, 349(1329):257–262, 1995.

[21] S. Gray, P. Szymanski, and M. Levine. Short-range repression permits multiple enhancers to function autonomously within a complex promoter. *Genes & Development*, 8(15):1829–1838, 1994.

[22] M. Gribskov, A.D. McLachlan, and D. Eisenberg. Profile Analysis: Detection of Distantly Related Proteins. *Proceedings of the National Academy of Sciences*, 84(13):4355–4358, 1987.

[23] V.V. Gursky, J. Jaeger, K.N. Kozlov, J. Reinitz, and A.M. Samsonov. Pattern formation and nuclear divisions are uncoupled in Drosophila segmentation: Comparison of spatially discrete and continuous models. *Physica D*, 197(3-4), 2004.

[24] V.V. Gursky, J. Reinitz, and A.M. Samsonov. How gap genes make their domains: An analytical study based on data driven approximations. *Chaos*, 11(1):132–141, 2001.

[25] M. Huelskamp, C. Pfeifle, and D. Tautz. A morphogenetic gradient of hunchback protein organizes the expression of the gap genes Krueppel and knirps in the early Drosophila embryo. *Nature*, 346(6284):577–580, 1990.

[26] Barkai N Ihmels J, Bergmann S. Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20:1993–2003, 2004.

[27] L. Ingber. Simulated annealing: Practice versus theory. *Mathl. Comput. Modelling*, 18(11):29–57, 1993.

[28] PW Ingham. The molecular genetics of embryonic pattern formation in Drosophila. *Nature*, 335(6185):25–34, 1988.

[29] J. Jaeger, M. Blagov, D. Kosman, K.N. Kozlov, E. Myasnikova, S. Surkova, C.E. Vanario-Alonso, M. Samsonova, D.H. Sharp, and J. Reinitz. Dynamical Analysis of Regulatory Interactions in the Gap Gene System of Drosophila melanogaster. *Genetics*, 167(4):1721–1737, 2004.

[30] J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K.N. KozlovManu, E. Myasnikova, C.E. Vanario-Alonso, M. Samsonova, and D.H. Sharp. Dynamic control of positional information in the early Drosophila embryo. *Nature*, 430:368–371, 2004.

[31] H. Janssens, S. Hou, J. Jaeger, A.R. Kim, E. Myasnikova, D. Sharp, and J. Reinitz. Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. *Nature Genetics*, 38:1159–1165, 2006.

[32] H. Janssens, D. Kosman, C.E. Vanario-Alonso, J. Jaeger, M. Samsonova, and J. Reinitz. A high-throughput method for quantifying gene expression data from early Drosophila embryos. *Development Genes and Evolution*, 215(7):374–381, 2005.

[33] SA Kauffman. Homeostasis and differentiation in random genetic control networks. *Nature*, 224(215):177–178, 1969.

[34] S. Kirkpatrick, CD Gelatt Jr, and MP Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671, 1983.

[35] D. Kosman, S. Small, and J. Reinitz. Rapid preparation of a panel of polyclonal antibodies to Drosophila segmentation proteins. *Development Genes and Evolution*, 208(5):290–294, 1998.

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

[37] CE Lawrence, SF Altschul, MS Boguski, JS Liu, AF Neuwald, and JC Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[38] H.H. Mc Adams and A. Arkin. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA*, 94:814–819, 1997.

[39] E. Mjolsness, DH Sharp, and J. Reinitz. A connectionist model of development. *J Theor Biol*, 152(4):429–53, 1991.

[40] E. Myasnikova, A. Samsonova, K. Kozlov, M. Samsonova, and J. Reinitz. Registration of the expression patterns of Drosophila segmentation genes by two independent methods. *Bioinformatics*, 17(1):3–12, 2001.

[41] E. Myasnikova, M. Samsonova, D. Kosman, and J. Reinitz. Removal of background signal from in situ data on the expression of segmentation genes in Drosophila. *Development Genes and Evolution*, 215(6):320–326, 2005.

[42] T.J. Perkins, J. Jaeger, J. Reinitz, and L. Glass. Reverse Engineering the Gap Gene Network of Drosophila melanogaster. *PLoS Comput Biol*, 2(5):e51, 2006.

[43] E. Poustelnikova. A database for management of gene expression data in situ. *Bioinformatics*, 20(14):2212–2221, 2004.

[44] GK Purushothama and L. Jenkins. Simulated annealing with local search-a hybrid algorithm for unit commitment. *Power Systems, IEEE Transactions on*, 18(1):273–278, 2003.

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

[46] J. Reinitz, S. Hou, and D.H. Sharp. Transcriptional Control in Drosophila. *Complexus*, 1(2):54–64, 2003.

[47] J. Reinitz, S. Hou, and D.H. Sharp. Known maternal gradients are not sufficient for the establishment of gap domains in Drosophila melanogaster. *Mechanisms of Development*, 124:108–128, 2007.

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

[49] DR Rigney. Stochastic models of cellular variability. *Kinetic Logic: A Boolean Approach to the Analysis of Complex Regulatory Systems*, 29:237–280, 1979.

[50] R. Rivera-Pomar and H. Jackle. From gradients to stripes in Drosophila embryogenesis: filling in the gaps. *Trends Genet*, 12(11):478–83, 1996.

[51] F.P. Roth, J.D. Hughes, P.W. Estep, and G.M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, 16(10):939–945, 1998.

[52] M. Rothe, EA Wimmer, MJ Pankratz, M. Gonzalez-Gaitan, and H. Jackle. Identical transacting factor requirement for knirps and knirps-related Gene expression in the anterior but not in the posterior region of the Drosophila embryo. *Mech Dev*, 46(3):169–81, 1994.

[53] L. Sánchez and D. Thieffry. A Logical Analysis of the Drosophila Gap-gene System. *Journal of Theoretical Biology*, 211(2):115–141, 2001.

[54] L. Sanchez, J. van Helden, and D. Thieffry. Establishement of the Dorso-ventral Pattern During Embryonic Development of Drosophila melanogaster: a Logical Analysis. *Journal of Theoretical Biology*, 189(4):377–389, 1997.

[55] M.D. Schroeder, M. Pearce, J. Fak, H. Fan, U. Unnerstall, E. Emberly, N. Rajewsky, E.D. Siggia, and U. Gaul. Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol*, 2(9):e271, 2004.

[56] E. Segal et al. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19(90001):273–282, 2003.

[57] AA Simcox and JH Sang. When does determination occur in Drosophila embryos? *Dev Biol*, 97(1):212–21, 1983.

[58] S. Sinha. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, 22(14):e454, 2006.

[59] S. Sinha et al. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19(90001):292–301, 2003.

[60] D. StJohnston and C. Nusslein-Volhard. The origin of pattern and polarity in the Drosophila embryo. *Cell*, 68(2):201–19, 1992.

[61] H. Szu. Fast simulated annealing. *AIP Conference Proceedings*, 151:420, 1986.

[62] RL Tatusov, SF Altschul, and EV Koonin. Detection of Conserved Segments in Proteins: Iterative Scanning of Sequence Databases With Alignment Blocks. *Proceedings of the National Academy of Sciences*, 91(25):12091–12095, 1994.

[63] S.R. Thangiah, I.H. Osman, and T. Sun. Hybrid Genetic Algorithm, Simulated Annealing and Tabu Search Methods for Vehicle Routing Problems with Time Windows. *Canterbury: University of Kent [cited 6.5. 1998]. Available from Internet:¡ URL: ftp://brutus. cpsc. sru. edu/pub/papers/GenSAT. ps. z*, 1994.

[64] R. Thomas. Boolean formalization of genetic control circuits. *J Theor Biol*, 42(3):563–85, 1973.

[65] P.J.M. van Laarhoven, E.H.L. Aarts, and J.K. Lenstra. Job Shop Scheduling by Simulated Annealing. *Operations Research*, 40(1):113–125, 1992.

[66] W. Wang, J.M. Cherry, Y. Nochomovitz, E. Jolly, D. Botstein, and H. Li. Inference of combinatorial regulation in yeast transcriptional networks: A case study of sporulation. *Proceedings of the National Academy of Sciences*, 102(6):1998–2003, 2005.

[67] B. Wilczynski and J. Tiuryn. Regulatory Network Reconstruction using Stochastic Logical Networks. *LNBI*, 42(10):145–157, 2006.

[68] C.H. Yuh, H. Bolouri, and E.H. Davidson. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development*, 128:617–629, 2001.