NEW INSIGHTS INTO CIS-REGULATORY MODULE EVOLUTION USING *IN-SILICO* EVOLUTIONARY SIMULATIONS

BY

THYAGO SELLMANN PINTO CESAR DUQUE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Associate Professor Saurabh Sinha, Chair
Professor Gerald F. DeJong, II
Assistant Professor Jian Ma
Associate Professor Alan Moses, University of Toronto

Abstract

Gene regulation is the process by which specific sets of genes are expressed in precise spatial/temporal patterns (Davidson 2010). It is a fundamental process with impact on development and cell identity (Fisher 2002; Davidson 2010), cancer (Riggs and Jones 1983; Ballestar and Esteller 2008) and other diseases such as Alzheimer's disease (van Duijn et al. 1999), and several other biological processes (Davidson 2010). Understanding gene regulation, and its evolution, is an important quest in biology and medicine, and it is one that is often addressed with the help of computational tools.

In this thesis we present a suite of computational tools and statistical methods developed to simulate the evolution of gene regulatory sequences in a realistic setting. We also describe new insights into function, mechanisms and evolution of gene regulation that have been learned with the help of these tools.

We first demonstrate the ability of our tools to model the evolution of regulatory sequences from 12 species of fruitflies. In our comparison with other available tools, we have been able to achieve better performances while using a smaller number of free parameters. Additionally, we describe three studies that provide new insights concerning the evolution and mechanism of the regulatory machinery.

As the first relevant insight, we demonstrate that the phenomenon of homotypic clustering of transcription factor binding sites, which is often associated with mechanistic implications or origins (e.g., cooperative activation), may also be explained as an evolutionary artifact, or, in the language of (Lusk and Eisen 2010), an evolutionary mirage.

Our second study demonstrates how the accurate modeling of evolutionary data for regulatory sequences can be used to elicit biophysical mechanisms of the regulatory machinery. Specifically, we demonstrate how discrepancies between our evolutionary model and real data pointed to a possible cooperative interaction between molecules of a transcription factor, which was then confirmed using biological essays.

Finally we use our tool to explore questions related to the time necessary to evolve an enhancer under a diverse set of situations. We find that some enhancers are easier to evolve than others and that a number of factors, including biophysical mechanisms and the starting point for evolution will impact the time necessary to evolve regulatory sequences.

The insights that we have been able to gain using our tools are relevant to biologists, but perhaps equally relevant is the fact that all these insights have been learned largely from one computational tool, which demonstrates the flexibility of our tool in particular, as well as the importance of computational biology approaches in general.

To mom, dad and Gaby

Acknowledgments

This project would not have been possible without the support of many people. Many thanks to my adviser, Saurabh Sinha, who at was patient and still pushed me to do my best work. Also thanks to my committee members, Gerald DeJong, Jian Ma, Caetano-Anolles and Alan Moses; and to my friends and lab mates, in special, Majid Kazemian, Hassan Samee, Charles Blatti III, Kai Zhao, Andrei Stefanescu, Traian Florin Şerbănuţă, Esteban Menezes-Rojas and Onur Peckan. Special thanks for my fiancée, Yukiko Kaneko, for her support and patience during the most stressful times and to my parents and Gaby for their unconditional support before, during and after.

I would also like to acknowledge the “Fundação de Amparo à Pesquisa do Estado de São Paulo”, FAPESP, for its financial support.

Table of Contents

1 Introduction

There has long been an effort to understand the regulatory logic involved in the regulation of gene expression. Efforts to this effect have involved a variety of tools and approaches, including sequence signature approaches (Sinha et al. 2003; Sinha et al. 2006; Kazemian et al. 2013), experimental approaches (Barrios-Rodiles et al. 2005; Vizoso Pinto et al. 2009; ENCODE Project Consortium 2011; Arnold et al. 2013; Kazemian et al. 2013), quantitative modeling (Jaeger et al. 2004; Segal et al. 2008; He et al. 2009; Fakhouri et al. 2010; He et al. 2010; Kaplan et al. 2011; Cheng et al. 2013), regulatory network approaches (Levine and Davidson 2005; Perkins et al. 2006; Zeitlinger et al. 2007) and evolutionary approaches (Moses et al. 2004; Moses et al. 2006; Francois et al. 2007; Kim et al. 2009; Lusk and Eisen 2010; Stewart et al. 2012).

While at first these approaches may seem diverse they often relate and complement each other. For example, the quantitative modeling approach often relies on existing experimental data and conversely experimental approaches often rely on modeling to attribute meaning to data. Sequence signature models often rely on evolutionary models to increase accuracy and distinguish relevant patterns and network approaches often include a quantitative component.

Evolutionary approaches in particular have received increasing attention lately, in part due the increased availability of evolutionary data (e.g. genomic sequences for multiple related species and population variation data) and in part due to the realization that evolutionary models can improve existing quantitative and regulatory network models as well as shed light on the interpretation of experimental data. For example, evolutionary models have been used to understand the circumstances under which a gene-regulatory system is likely to display certain properties (e.g. cooperative gene regulation (Stewart et al. 2012) or specific network architectures (Cooper et al. 2009)). Conversely, assumptions pertaining to the evolution of gene-regulatory systems have been use to improve sequence signature models (e.g. Stubb (Sinha et al. 2003; Sinha et al. 2006) and iTFs (Kazemian et al. 2013))) or to offer a null hypothesis against which such sequence signature models should be judged (Lusk and Eisen 2010).

Additionally, biologists are often interested in the evolutionary process itself, with important questions relating to 1) how long does it take for specific regulatory sequences to evolve under certain assumptions (Stone and Wray 2001; Carter and Wagner 2002; Gerland and Hwa 2002; MacArthur and Brookfield 2004; Durrett and Schmidt 2007, 2008), 2) whether certain regulatory sequences display signatures of positive selection or negative selection (Moses et al. 2004; Moses 2009; He et al. 2011), 3) what is the evolutionary history of certain regulatory sequences (Francois et al. 2007; Josephides and Moses 2011), 4) how to best model the evolution of regulatory sequences (Berg et al. 2004b; Kim et al. 2009; Nourmohammad and Lässig 2011); among many others.; among many others.

In this thesis we propose a suite of computational tools and statistical methods designed to help us model the evolution of gene regulation. Our tool is designed to be as generic and realistic as possible, and therefore, to be able to address many of the questions introduced in the previous paragraphs. Using the proposed tools, we have been able to acquire new insights into the mechanisms and evolution of gene regulation.

At the center of our suite of tools is a simulation software called Predicted-Expression-Based CRM Evolution Simulator (PEBCRES), which is a Wright-Fisher (Hartl and Clark 1997) simulation framework for (cis-Regulatory Modules) CRMs. It uses GEMSTAT (He et al. 2010), a realistic sequence-to-expression model, to estimate the fitness of evolving regulatory sequences based on their expression.

Using PEBCRES we have been able to achieve important insights related to four biologically relevant objectives, namely (1) modeling evolutionary data; (2) detecting evolutionary artifacts; (3) understanding the mechanisms of gene regulation and (4) understanding the evolution of regulatory sequences. These objectives will be introduced bellow and discussed in depth on Chapters 4 to 7.

1.1 Objective 1: Modeling evolutionary data

Description: The first step in this project is to create a simulation framework that can realistically model the evolution of regulatory sequences. The term “regulatory sequences”

here refers mainly to “enhancers” or “cis-Regulatory Modules” (CRMs), which are described in the next section.

Approach: We created a Wright-Fisher simulation framework (Hartl and Clark 1997) that uses a realistic sequence-to-expression model in assigning fitness to regulatory sequences (CRMs) based on their expression. The sequence-to-expression model is based on a statistical thermodynamics model (Shea and Ackers 1985; Buchler et al. 2003; Segal et al. 2008; He et al. 2010) and provides a realistic way of assigning fitness to any given sequence , typically an evolving CRM.

Impact: An accurate and realistic model of regulatory sequence evolution can be used to predict the evolutionary fate of sequences and as a baseline for what is expected under certain assumptions. Such predictions can then be compared to evolutionary patterns gleaned from orthologous regulatory sequences, thus providing a model for evolutionary data. The model can be used to test hypotheses regarding the evolutionary process and biophysical properties of the transcription process, enabling aims 2-4.

1.2 Objective 2: Detecting evolutionary artifacts

Description: Evolutionary artifacts are sequence patterns derived from the evolutionary process that could be misinterpreted as functionally relevant (having mechanistic importance) (Lynch 2007b; Lusk and Eisen 2010). We use our simulation framework to detect such phenomena by establishing a realistic baseline (null model) to test hypotheses against.

Approach: We use our model to evolve sequences under adaptive selection and observe baseline characteristics of resulting sequences, for example, finding typical site counts and variance thereof. We also use analytical models to complement the analysis.

Impact: Several methods use sequence signature and statistical analysis to infer CRMSs or to generate mechanistic hypotheses about transcriptional regulation. We provide cautionary measures to use when evaluating such hypotheses purely through data analysis. This can

reduce false positive in CRM prediction or the risk of misclassifying an evolutionary artifact as a mechanistic feature.

1.3 Objective 3: Understanding the mechanisms of gene regulation

Description: The underlying mechanistic features of the transcription process, which are often hard to elicit through single species modeling only, can be made clearer using evolutionary data (as reported in [Duque et al. 2013 (in revision)]). The objective is finding evidence for biophysical mechanisms involved in the transcription process, e.g., interactions among relevant players in gene regulation.

Approach: A hypothesis about the underlying mechanism of transcription is tested by simulating evolutionary data under a null model that excludes the mechanism and under an alternative model that incorporates it. Conclusions about the validity of the hypothesis are drawn based on the agreement with evolutionary data from each model, and the hypothesis can be later tested *in vitro* or *in vivo* (with the decision to test being guided by the results from our approach).

Impact: Testing a hypothesis about a mechanism of interaction in the transcription process is expensive and time consuming. Our approach can elicit the most promising features and help prioritize such tests.

1.4 Objective 4: Understanding the evolution of regulatory sequences

Description: There are many characteristics intrinsic to the evolutionary processes and to specific expression patterns that we would like to understand better. These include questions like “how hard is it to evolve is a specific expression pattern and why?”, “what features improve or impede the evolvability of a sequence or pattern?” and “what is the effect of positive and negative selection in regulatory sequences?”

Approach: We use evolutionary simulations to evolve a regulatory sequence that drives a specific expression pattern, starting from a random sequence or from a sequence associated

with a different pattern. We study how different models (both evolutionary and of transcription) affect the time necessary to evolve a CRM that drives that target expression.

Impact: To the best of our knowledge, there is no other evolutionary model capable of predicting the time necessary for complete CRMs to evolve. However, PEBCRES can be used to produce such estimates under a variety of assumptions. Additionally, understanding how different features help or hinder evolution can help understand what evolutionary forces lead to the current regulatory programs observed in nature.

1.5 Relevance and impact

A deep knowledge of the mechanisms of gene regulation is essential to understand many biological processes, including development (Fisher 2002; Davidson 2010) and disease (Riggs and Jones 1983; van Duijn et al. 1999). Important questions relating to the mechanisms of gene regulation include what role specific TFs play in determining the level and boundaries of gene expression, which TFs interact with each other and in what ways, whether a repressor acts as a short range repressor or as direct repressor, what is the importance of shadow enhancers, among many others.

Many of these questions can be and are traditionally answered with direct experimental assays. However, these experiments are often expensive to perform and an unguided search for all possible mechanistic features is simply impractical. Computational modeling and evolutionary data are two useful tools to prioritize further experimental investigation. Examples of works in which computational modeling is used to suggest or test hypotheses regarding gene regulation mechanisms include (Perkins et al. 2006; He et al. 2010; Parker et al. 2011; Kanodia et al. 2012). In this work we focus on the use of evolutionary data and simulation for mechanistic inference.

Evolutionary data can be used to infer the workings of several of the mechanisms of the regulatory process. For example, Hare et al. (2008) studied the *even-skipped* locus at six species of scavenger fruitflies. These species are highly diverged from *D. melanogaster* and display little sequence similarity despite producing matching expression pattern. However, they found a number of nearly perfectly conserved short (20-30 bp) sequences, which were strongly

enriched in pairs of binding sites, either overlapping or adjacent. They hypothesized that the particular local arrangement of those sites relative to each other was more relevant than their global arrangement within the CRM.

Similarly, Kim et al. (2009) found a number of statistically significant patterns that could be interpreted as indication for assorted mechanistic features, though they also caution against taking these as a fact. Examples of statistical patterns observed included the reduced likelihood of loss for sites of some TFs when adjacent to another site of the same factor (which can be taken to indicate direct or indirect interaction) and the effect of a proximal or overlapping site on a site's evolution (which can be interpreted in similar manner as (Hare et al. 2008))

However, despite the feasibility and promise of inferring mechanistic features from evolutionary data, Lusk and Eisen (2010) caution about evolutionary artifacts (or mirages in their terminology) that may lead to false inference. They notice that the deletion bias (Tanay and Siggia 2008) in *D. melanogaster* may lead to an enrichment of proximal and overlapping sites even in the absence of any mechanistic feature related to these properties. They reach this conclusion using a simple evolutionary simulation. Their results not only caution us again false inferences but also showcase the power of evolutionary simulations, or *in silico* evolution, to clarify (or refute) particular mechanisms of gene regulation. The use of evolutionary simulation to clarify aspects of gene regulation (or biology in general) need not restricted to mechanistic features. Francois et al. (2007) use *in silico* evolution to learn about the evolution of segmentation in insects, concluding, for example, that the inter-conversion between short germ and long germ modes of development may have occurred multiple times. However, both Lusk and Eisen (Lusk and Eisen 2010) and François et al. (Francois et al. 2007) designed simulation frameworks to answer only the specific questions they were interested in. A number of other evolutionary models including (Cooper et al. 2009; Josephides and Moses 2011; Stewart et al. 2012) also tend to be designed for very specific scenarios. Inspired by these previous efforts, our goal here was to develop a general purpose model of regulatory sequence evolution to answer a wide range of questions about the evolution and mechanisms of gene regulation.

1.5.1 Common issues of models of regulatory sequence evolution

A special class of models of regulatory evolution includes the models by Lässig and co-workers (Berg et al. 2004b; Mustonen and Lässig 2005) and the model by Kim et al. (2009), as well as the models by Stone and Wray (2001), Durrett and Schmidt (2007, 2008). These models differ from the models by (Francois et al. 2007; Cooper et al. 2009; Josephides and Moses 2011; Stewart et al. 2012) in an important aspect: they simulate evolution at the level of regulatory sequence, rather than at, say, the level of a regulatory network with changing nodes and edges. Particularly, the models by Lässig and co-workers (Berg et al. 2004b; Mustonen and Lässig 2005) and Kim et al. (2009) assume that the fitness of a sequence depends on the binding energy of its sites, which in turn is assumed to relate to the PWM score of the site. However these models typically focus on single binding sites and consequently neither model is capable of perfectly matching the evolutionary data they model, leaving room for improvement. For example, the Site-level Simulator (SS) model from Kim et al. (2009) fits the observed patterns of binding site conservation between *D. melanogaster* and *D. yakuba* better than the Halpern-Bruno model does. However, the SS model still predicts an under-conservation of sites when compared to the data from CRMs in *D. melanogaster* and *D. yakuba.*

We speculate that there are at least four reasons why most models of evolution for regulatory sequences fail to accurately model real evolutionary data (hereby referred as Common Issues 1-4):

1) *Continuous nature of the functional effect of mutations*. As mentioned above, the SS model assumes that the functional effect of a site is determined by whether the site is strong or weak. In other words, the SS model defines a binary fitness for binding sites. However, it is reasonable to expect that the functional contribution of a binding site can potentially be at multiple levels based on the binding energy of the site (Stormo and Fields 1998), which is typically estimated by the agreement between the sequence and the TF motif.
2) *Context in which a binding site evolves*. The SS model assumes that each site in a CRM evolves independently of all other sites in that CRM, and in a manner that is independent of the expression driven by the CRM, i.e., independent of CRM function. In reality, however,

one may expect that some binding sites can tolerate deleterious changes (Spivakov et al. 2012) while other sites stay immutable across large evolutionary distances (Visel et al. 2008), even if the sites are bound by the same TF and with similar strengths. This is because the fitness consequence of an in-site mutation, and hence its evolutionary fate, depends on the contribution of that site to the CRM's regulatory function, and the precise effect of the mutation on function. The SS model tries to mitigate this issue to an extent by learning a different selection coefficient for each transcription factor. This approach only captures the fact that sites from different factors evolve differently, while forcing sites from the same factor to evolve under the same constraints, regardless of context. Here, context may refer to the entire CRM or to the immediate neighborhood of a site. For instance, a given CRM may have a functional excess of sites for a specific TF, thereby reducing the selective pressure for individual sites. On the other hand, it is also possible that a nearby site increases the selective pressure by mediating cooperative or competitive binding.

3) *Evolutionary changes in the context of the binding site*. Since the SS model (as well as the models by Lässig and co-workers (Berg et al. 2004b; Mustonen and Lässig 2005)) ignores the context of a site, it also ignores evolutionary changes in the context. In reality, the context of a site evolves with the site and may lead to interesting evolutionary dynamics, such as compensatory mutations in two different sites of the same CRM (Ludwig and Kreitman 1995; Carter and Wagner 2002; Durrett and Schmidt 2008). For example, the strong pressure for conservation of a site could be relaxed if a nearby site from the same TF is made stronger or if a new site for the same TF is created. Conversely, a site under relatively weak pressure for conservation could be forced into a situation where no mutations are tolerated, by the weakening of a nearby site of the same TF or by the strengthening of a site with an opposing regulatory effect.
4) *Combinatorial regulation by multiple TFs*. Real metazoan CRMs are usually composed of several binding sites for several TFs which act in combinatorial ways to create intricate spatial-temporal patterns. For example, the Anterior-Posterior patterning system in Drosophila's *blastoderm* stage sets up complex multiple-stripe patterns for genes like *eve, hairy* and *runt* via a complex combinatorial network of TFs including maternal deposited TFs

(e.g. *Bicoid*), TFs produced by the "gap" genes (e.g. *Giant*) and TFs expressed in the terminal ends of the embryo (e.g. *Huckebien*) (St Johnston and Nusslein-Volhard 1992; Furriols and Casanova 2003). These TFs have a variety of roles (e.g. activation by *Bicoid,* repression by *Giant*, chromatin remodeling by *Zelda* (Harrison et al. 2011)) and a variety of expression patterns (e.g. anterior expression of *Bicoid*, uniform expression of DSTAT, stripped pattern of *Giant*) that are combined to form intricate and precise spatial patterns. Without a model capable of capturing the combinatorial regulation by multiple TFs we are unlikely to fully understand the complexity of the regulatory machinery and its evolution.

These issues are not limited to the SS model, but rather are common to many models of regulatory sequence evolution. For example, the models by Stone and Wray (2001) and Durrett and Schmidt (2007, 2008) were designed with a different for an objective different from the models by Lässig and co-workers (Berg et al. 2004b; Mustonen and Lässig 2005) and Kim et al. (2009); they were designed to estimate the time necessary for regulatory sequences to evolve under certain assumptions. Still, both Stone and Wray (2001) and Durrett and Schmidt (2007) suffer from common issues 1-4 since they model evolution at the level of a single binding site. These models are, therefore, limited to insights at the binding site level, being unable to provide much insight into the evolution of complete CRMs.

There are, however, models that address at least partially one or more of these issues. For example, (Carter and Wagner 2002) and (Durrett and Schmidt 2008) partially address common issues 2 and 3 by modeling the evolution of a pair of binding sites instead of a single binding site. While these models would not be able to explore many important issues (e.g. homotypic clustering, see Chapter 4) since real CRMs in metazoans are usually composed of many binding sites, the models by (Carter and Wagner 2002) and (Durrett and Schmidt 2008) demonstrate that deeper biological insights can be obtained when modeling (even if only partially) the dependent evolution of multiple binding sites. For example Carter and Wagner (2002) demonstrated the importance of the phenomenon that was latter called stochastic tunneling (Iwasa et al. 2004), complementing the previously accepted explanation for binding site

turnover events and explaining why these events are more common in invertebrate populations than in vertebrate populations.

The model by MacArthur and Brookfield (2004) is probably the closest from addressing all of the mentioned common issues, modeling the evolution of sets of binding sites for a single TF. However, their model is not cable of modeling the combinatorial regulation from multiple TFs (common issue 4) and suffers from other methodological issues (see Chapter 7), limiting the insights that can be learned from the model.

In this thesis we address all of the common issues mentioned above via a new framework for *in-silico* evolution, based on a state-of-the-art sequence-to-expression model. Our model can accurately fit evolutionary data (objective 1) and is capable of discerning evolutionary artifacts (objective 2) and mechanistic features (objective 3). Additionally, our model can be used to understand the how evolution acts on regulatory sequences (objective 4).

The remainder of this thesis is organized in the following way: Chapter 2 briefly introduces the basic background necessary to follow this thesis, including the basic principles of gene regulation and evolutionary models. Chapter 3 expands on Chapter 2 and describes in larger detail the tools and methods proposed in this thesis. Chapter 4 describes the performance of our tools on the task of modeling evolutionary data from 12 Drosophila species. Chapters 5, 6 and 7 describe three important biological insights obtained using our tool. Finally, Chapter 8 concludes this thesis as summarizes our work.

2 Background

In this chapter we introduce the basic background necessary to follow this thesis, introducing the basic terminologies used throughout this thesis and listing a few of the basic references for additional information.

Basic principles of gene regulation: The precise spatial/temporal expression of genes is controlled by proteins called Transcription Factors (TFs). These specialized proteins bind to the DNA and interact with the transcription machinery helping activate or repress the transcription process. A TF usually binds to short sequences of DNA called Transcription Factor Binding Sites (TFBS) or simply binding sites.

A stretch of DNA that harbors a relatively large number of binding sites and is functionally implicated in transcription is called a *Cis*-Regulatory Module (CRM) or enhancer (Davidson 2010). In *Drosophila,* these sequences typically range from 500 to 2000 base pairs (bp) and are located 5' to the Transcription Start Site (TSS) (Li et al. 2007a), but can also be found on other areas including 3', introns and even in the intergenic region of other genes (Perry et al. 2010; Barolo 2012).

The binding specificity of a TF is usually modeled using a Position-specific Weight Matrix (PWM) (Stormo and Fields 1998), which stores, for each position in the binding site, the probability that the TF will bind to each base in that position. The log-likelihood ratio (LLR) of a site computed according to the PWM (also referred as PWM score) can be interpreted as the binding energy of the TF for that specific site (Stormo and Fields 1998).

Mechanistic Features: Throughout this proposal we use the terms mechanistic features or underlying mechanisms to denote any biochemical process that influences the transcription of a gene. An example of a mechanistic feature is DNA- binding by a TF or activation by a TF; however, many other mechanisms are relevant to gene regulation. Some of the mechanisms of interest for this work include (non-exhaustive list):

Cooperative interaction: The mechanism by which one factor influences, through direct interaction or an indirect mechanism, some property of another factor. Examples of properties that can be influenced include DNA binding and/or activation strength (Giniger and Ptashne 1988; Hertel et al. 1997; Burz et al. 1998; Staten et al. 2004; Lebrecht et al. 2005).

Short range repression: The mechanism by which a transcription factor can repress the normal function of another TF via proximally located binding sites (Gray and Levine 1996).

Synergistic activation: The mechanism by which several different TFs can interact simultaneously with the Basal Transcription Machinery (BTM). The degree of synergistic activation (or the number of players that can interact that the same time with the BTM) is an important feature of the gene regulation process (Lin et al. 1990; Joung et al. 1993; He et al. 2010).

Indirect activation: The mechanism by which a transcription factor increases the transcriptional output or BTM occupancy through interaction with another (bound) transcription factor but without directly interacting with the BTM (Kanodia et al. 2012).

Chromatin remodeling: The mechanism by which so-called "pioneer factors" (Harrison et al. 2011; Nien et al. 2011) change the local chromatin structure thereby influencing DNA-binding by other TFs.

Shadow enhancers: Distal CRMs that help to drive specific expression patterns by acting in addition to a more gene-proximal CRM that drives the same pattern. These CRMs can help establish precise patterns or reinforce existing ones (Perry et al. 2010; Barolo 2012).

Anterior/Posterior patterning in Drosophila: The anterior/posterior patterning system (AP system) (Meinhardt 1978; Akam 1987) in *D. melanogaster* embryos is one of the best understood regulatory systems. Acting during the *blastoderm* stage of development, this system sets up a stripe pattern for genes including *eve, hairy* and *runt*. These patterns are set by a complex network that includes maternal deposited TFs (e.g. *Bicoid*), a set of "gap" genes

(e.g. *Giant*) and a set of terminal genes (e.g. *Huckebien*) (St Johnston and Nusslein-Volhard 1992; Furriols and Casanova 2003).

The AP system is ideal to study gene regulation since we have (1) experimentally characterized binding motifs for the relevant TFs (Noyes et al. 2008) (2) experimental knowledge of the relative concentration of these TFs (Tomancak et al. 2002; Tomancak et al. 2007), (3) *in-vivo* expression profiles for the target genes (Tomancak et al. 2002; Tomancak et al. 2007; Segal et al. 2008) and (4) experimentally validated CRMs (Halfon et al. 2008a). Moreover, there are 12 *Drosophila* genomes available, spanning from very short (e.g. about 2.5 million years between *D. melanogaster* and *D. simulans* (Ranz et al. 2003)) to relatively longer (e.g. about 65 million years between *D. melanogaster* and *D. mojavensis* (Staten et al. 2004)) evolutionary distances. For these reasons we chose the AP system for this study.

Evolutionary models: The evolution of a DNA sequence can be modeled by various broadly applicable evolutionary models (Kimura 1980; Felsenstein 1981; Hasegawa et al. 1985; Halpern and Bruno 1998). These are relatively simple evolutionary models that can often be directly applied to several domains. However, due to their general-purpose nature, they often fail to capture specific evolutionary constraints when modeling special types of sequences, such as regulatory sequences.

On the other hand, there also exist evolutionary models that are designed for a specific type of sequence (e.g., protein coding sequences (Halpern and Bruno 1998), structured RNA (Bradley and Holmes 2009) , etc.). These take into consideration specific characteristics of the sequences they are modeling and provide better fits to observed evolutionary data from that domain.

In recent years, there has been growing interest in evolutionary models for regulatory sequences such as transcription factor (TF) binding sites, especially in light of reports of frequent binding site turnover despite functional constraints (Moses et al. 2006; Doniger, Fay 2007; Spivakov et al. 2012) and also reports of ultra-conserved genomic segments being associated with regulatory function (Visel et al. 2008).

Evolutionary models designed specifically for regulatory sequences include the models designed by Lässig and co-workers (Berg et al. 2004b; Mustonen and Lässig 2005). These evolutionary models assume selection to act on the entire binding site rather than position by position as it is common in generic models. The same idea was used by Kim et al. (Kim et al. 2009) to model TFBS evolution in *Drosophila* CRMs. When compared to the broadly applicable evolutionary models, the models from Kim et al. and from Lässig and coworkers largely improve the agreement between model prediction and observed data, as is usually the case for application-specific models.

Additionally, a number of evolutionary models have been used to study specific characteristics of the regulatory evolution process without explicitly modeling sequence evolution. These include population genetics models (Stewart et al. 2012), network level evolution models (Cooper et al. 2009; Pujato et al. 2013) and explicit modeling of all likely evolutionary path (Josephides and Moses 2011).

Computational models of Gene Expression: He et al. (He et al. 2010) developed the GEMSTAT model, a model to predict the expression pattern induced by a regulatory sequence, using a statistical thermodynamics approach. GEMSTAT is able to accurately model the expression pattern of 37 CRMs in the AP system making it an ideal model of gene expression (1) accurately and realistically models real patterns, (2) allows for the implementation of several mechanistic features including short and long range repression (Gray and Levine 1996), cooperative DNA binding (Giniger and Ptashne 1988; Hertel et al. 1997), indirect activation (Kanodia et al. 2012), synergistic activation (Lin et al. 1990; Joung et al. 1993; He et al. 2010), etc. and (3) is computationally efficient for the task of predicting the expression of a sequence, which involves a dynamic programming implementation.

GEMSTAT has recently been extended to model the entire gene locus (Samee et al. 2013 (in revision)) and to model other patterning systems including the DV system (M.A.H. Samee, personal communication). Other models of gene expression include (Jaeger et al. 2004; Zinzen et al. 2006; Segal et al. 2008; Fakhouri et al. 2010; Kim et al. 2013; Ostuni et al. 2013).

Since GEMSTAT is the sequence-to-expression model used in our tools, we will describe it in further details in Chapter 3.

3 The PEBCRES simulation software

In this chapter we describe the PEBCRES simulation software, designed to simulate the evolution of regulatory sequences in a realistic, efficient and flexible manner.

The PEBCRES model can be divided into three components: (1) a sequence-to-expression model, that translates a given CRM sequence to the spatio-temporal expression pattern it will drive (2) a fitness function based on the CRM's predicted expression readout, and (3) an evolutionary simulation model. The evolutionary model is responsible for generating mutations and simulating the evolutionary fate of those mutations. The fate of each mutation is determined in part by the fitness of the mutated sequence, calculated via the fitness function, which is in turn based on a comparison between the expression pattern predicted for the mutated sequence and a target expression pattern. The sequence-to-expression model is used to predict the expression pattern driven by the sequence, enabling the comparison with the target pattern. In the following sections we will describe each of the components of the PEBCRES and finally, on Section 3.4, we will describe how these components are combined in our model.

3.1 The sequence-to-expression model: GEMSTAT

GEMSTAT is a sequence-to-expression model designed by He et al. (2010) to predict, across cell types, the level of gene expression driven by a CRM, given its sequence and information regarding the relevant TFs.

The expression pattern of a CRM is defined as the level of gene expression driven by that CRM across a set of cell-types. For example, the expression pattern of a CRM could be the level of expression driven by that CRM across the AP axis in the *Drosophila* embryo at the *blastoderm* stage (see Chapter 2), representing a spatial pattern. Another example is the level of expression driven by the CRM in a single cell, across different time-points. In the latter case the expression

[*] This chapter includes material previously published in (He et al. 2010; He et al. 2012; Duque et al. 2013)

pattern could represent the temporal changes to a cell as it changes from a pluripotent cell-type to increasingly specialized cell-types until it finally settles into a fully specialized cell-type.

The GEMSTAT model can predict the expression pattern for a CRM given the following input: (1) the binding specificity (PWM) of the relevant TFs; (2) the level of expression for each of the TFs in each of the cell-types; and (3) the sequence of the CRM. The inputs for GEMSTAT are generally be obtained experimentally or downloaded from existing databases (e.g. (Tomancak et al. 2002; Tomancak et al. 2007; Bryne et al. 2008; Halfon et al. 2008a; Noyes et al. 2008; Wingender 2008; Marygold et al. 2013; Mathelier et al. 2013)).

Additionally, the GEMSTAT model needs a series of parameters representing biophysical properties of the TFs and of the BTM, including at least one global parameter, the basal transcription level, and two parameters for each TF k:, the DNA binding parameter (β_k) and the transcriptional effect parameter (α_k). GEMSTAT also offers the option of modeling cooperative interactions between any two molecules of a TF. These parameters have to be obtained by training a GEMSTAT model on real CRMs with known expression patterns (for example from in-situ hybridization (Tomancak et al. 2002; Tomancak et al. 2007)) using the procedure from He et al. (2010) described on section 3.1.2. The meaning of each of these parameters is explained bellow.

3.1.1 Estimating expression from sequence

The GEMSTAT model uses concepts from statistical thermodynamics and an efficient dynamic programming implementation to estimate the level of expression driven by a CRM in each cell-type. Their statistical thermodynamics model follows (Buchler et al. 2003), dividing the model in two components: one dealing with the occupancy of TFs, and the another dealing with the interactions between occupied TFs and the BTM. TF occupancy influences the strength of TF-BTM interactions, which in turn determines gene expression levels.

For the first component, dealing with TF occupancy, we endeavor to calculate the relative probability $P(\sigma)$ of a configuration σ, which specifies which binding sites are bound and which binding sites are free (notice that for a CRM with n binding sites, there are 2^n possible

configurations). The probability $\mathrm{P}(\sigma)$ can be calculated as $\mathrm{P}(\sigma) = \frac{W(\sigma)}{Z}$, where $W(\sigma)$ is the statistical weight of configuration σ and $Z = \Sigma_{\sigma} W(\sigma)$ is the partition function.

The second component deals with how a given configuration σ affects gene expression. GEMSTAT assumes that the gene expression level is proportional to the fractional occupancy of the BTM (denoted as *E*). To calculate this quantity, we now consider that each of the 2^n configurations σ mentioned above can correspond to two states, one in which the BTM is bound and one in which the BTM is not. The statistical weight of the bound state is given by $W(\sigma)Q(\sigma)$ while the statistical weight of the unbound state is given by $W(\sigma)$. Here $Q(\sigma)$ is a contribution from TF-BTM interactions. The fractional occupancy can now be calculated as:

$$E = \frac{Z_{ON}}{Z_{OFF}} = \frac{\Sigma_{\sigma} W(\sigma)Q(\sigma)}{\Sigma_{\sigma} W(\sigma)Q(\sigma) + \Sigma_{\sigma} W(\sigma)} \qquad (1 **)$$

where $Z_{ON} = \Sigma_{\sigma} W(\sigma)Q(\sigma)$ is the relative probability of BTM being bound and $Z_{OFF} = \Sigma_{\sigma} W(\sigma)$ is the relative probability of the BTM being unbound.

The actual formulations of $W(\sigma)$ and $Q(\sigma)$ depend on specific mechanistic assumptions regarding the regulatory machinery, for example, these quantities depend on whether the model includes short-range repression, synergistic effects, cooperative binding, etc. In this thesis we will only describe one of the formulations, and the associated dynamic programming procedure for efficient computation of E. Information on the alternative formulations can be found on the original publication (He et al. 2010).

The Direct Interaction Model:

The direct interaction model is one of the models available in GEMSTAT. It reflects the assumption that bound TFs interact directly with the BTM, either favorably (activators) or unfavorably (repressors), thus affecting the probabilities that the BTM is bound, and consequently, the gene expression level. An alternative formulation is one in which only activators interact directly with the BTM while repressors block the effect of activators bound close to the repressor binding site. (This latter formulation is not pursued here.)

In the direct interaction model the contribution $q(S)$ of a single occupied binding site S to $W(\sigma)$ is given by:

$$q(S) = \beta_k [TF_k] e^{LLR(S) - LLR(S_{max})}$$

where β_k is a TF-specific free parameter[1] (representing the DNA binding parameter for TF k), $[TF_k]$ is the concentration of TF k, $LLR(\cdot)$ is the log-likelihood score of a site, calculated based on k's PWM and S_{max} is consensus site for k. Notice that GEMSTAT makes the assumption that the contribution of each position of the site is additive, an assumption that seems reasonable for the AP system (Janssens et al. 2006; Segal et al. 2008), but whose generality has been questioned (Benos et al. 2002; Zinzen et al. 2006).

We now can calculate the statistical weight of a configuration σ as:

$$W(\sigma) = \prod_i q(S_i)^{\sigma_i} \times \prod_{(i,j)|i<j} \omega_{ij}^{\sigma_i \sigma_j}$$

where the first product represents the individual contribution of each binding site to the statistical weight and σ_i is an indicator variable for whether site S_i is bound in configuration σ. The second product represents the TF-TF interaction contribution to $W(\sigma)$, with ω_{ij} representing the contribution from interactions between sites S_i and S_j. For our experiments $\omega_{ij} \neq 1$ only if S_i and S_j are sites for the same TF, S_i and S_j are adjacent in the configuration σ (i.e., there are no sites between S_i and S_j or the sites in between are not bound) and S_i and S_j are within 50 bp of each other. In that case ω_{ij} is a TF specific free parameter $\omega_k \geq 1$, where k indexes the TF bound to sites i and j. Notice that $\omega_{ij} = 1$ means there is no contribution to $W(\sigma)$ from TF-TF interaction for sites S_i and S_j.

We can also calculate the contribution to the statistical weight from the TF-BTM interactions as:

[1] Notice that in (He et al. 2010) the free parameter β is written as $K(S_{max})v$.

$$Q(\sigma) = \prod_i \alpha_{k_i}^{\sigma_i}$$

where σ_i an indicator variable as above and α_{k_i} is a TF-specific parameter representing the transcriptional effect of the TF k_i bound to site S_i in configuration σ.

Efficient implementation of the GEMSTAT model:

We now approach the question of how to efficiently compute the fractional occupancy of the BTM (E). Notice that this computation requires a summation over an exponential number of configurations and therefore the naive approach to this calculation would be intractable for real CRMs. Instead, He et al. (2010) proposed an efficient dynamic programming algorithm to compute this quantity. To that effect, we first calculate, for every site i, $Z_{OFF}(i)$, representing the total statistical weight summed over all configurations of sites up to site i, with site i being occupied ($\sigma_i = 1$). The recurrence to calculate $Z_{OFF}(i)$ is given by:

$$Z_{OFF}(i) = q(i)\left[\sum_{j \in \Phi(i)} \left[\omega_{i,j}\, Z_{OFF}(j)\right] + 1\right]$$

where $q(i)$ is the statistical weight of site i defined above, $\omega_{i,j}$ indicates the interaction between occupied sites i and j (see above) and $\Phi(i)$ is the set of all sites to the left of i that do not overlap i. Z_{OFF} can be trivially calculated as $\sum_i Z_{off}(i)$.

Similarly, we can calculate $Z_{ON}(i)$, defined as the analogous of $Z_{OFF}(i)$, as:

$$Z_{ON}(i) = q(i)\alpha_{k_i}\left[\sum_{j \in \Phi(i)} \left[\omega_{i,j} Z_{ON}(j)\right] + 1\right]$$

where k_i represents the TF bound to site i and α_{k_i} is the transcriptional effect of TF k_i (defined above). Similarly, we can calculate $Z_{ON} = \sum_i Z_{ON}(i)$.

The computational complexity of the calculation of Z_{ON} is $O(n^2)$ (n is the number of sites). This is due to the need to calculate $Z_{ON}(i)$ for every site i, which in turn involves the summation of

every site j to the left of i. The calculation of Z_{OFF} has the same complexity, and therefore the complexity of calculating E is $O(n^2)$. Additionally, the procedure to compute E has to be repeated once for every cell-type, since each cell-type has different concentration levels ($[TF_k]$) for each TF resulting in different values of $q(S)$ and, consequently, E.

We can now formally define the predicted expression pattern for a CRM, denoted as a vector $v = [v_1 \ v_2 \ \dots \ v_{nc}]$ where v_i represents the fractional occupancy (E) of the BTM (calculated as outlined above) on cell-type $i \in [1 \,..\, nc]$, with nc representing the number of cell-types (typically 60 in our experiments). This profile is used to calculate the fitness of arbitrary sequences in PEBCRES (see section 3.2).

Binding site calling:

Binding site calling is a problem in many TF-based models of gene regulation since it is difficult to express the intrinsically stochastic binding site occupancy by a TF as simple a binary variable (representing whether a subsequence is a binding site or not). Approaches to binding site calling often amount to estimating whether the binding site is expected to be occupied with probability above some threshold, but deciding on the threshold is often difficult.

The statistical thermodynamics approach of GEMSTAT allows us to ignore the problem of binding site calling since the model meaningfully weights the contributions of strong and weak binding sites. We could therefore call any sequence a binding site for any TF and rely on the fact that the contribution of weak sites to the statistical weight would be negligible. However even the efficient implementation of the GEMSTAT model requires computation that scales quadratically with the number of binding sites; hence, calling every sequence a binding site is likely to lead to poor run times.

As a balance between a model that fully weights the contributions to the statistical weight of every subsequence and a model that is computationally fast and efficient, we choose an very weak threshold on binding sites. This threshold, set to 4.0, is significantly smaller than the thresholds typically used on models of gene regulation, excluding only the truly irrelevant subsequences from the statistical weight computations while allowing for fast predictions.

3.1.2 Model training

As mentioned in the previous section, GEMSTAT has a series of TF-specific parameters that need to be estimated from data. The parameters are, for each TF k, the DNA-binding parameter β_k, the transcriptional effect parameter α_k, and for select TFs the self-cooperativity parameter ω_k. (Notice that GEMSTAT allows for the modeling of TF-TF interactions between any pair of TFs; however, in this thesis we choose not to include interactions between different TFs, due to the lack of statistical support for the inclusion of extra free parameters and the associated risk of overfitting.)

To train the free parameters of GEMSTAT, we use a set of 37 CRMs previously shown to drive patterned expression along the anterior-posterior (A/P) axis in the *blastoderm* stage *D. melanogaster* embryo. The expression pattern of each of these CRMs, determined experimentally, is represented as a 60-dimensional (i.e. $nc = 60$) vector of values in the range [0,1], with the dimensions of the vector corresponding to uniformly spaced positions along the AP axis from 20% egg length to 80% egg length. CRM sequences and their experimental expression profiles were collected by He et al. (2010). The relevant TFs used to model CRM function were BCD, CAD, KR, KNI, GT and HB. TF motifs were taken from the Fly Factor Survey database (Noyes et al. 2008).

This dataset is used to learn the values of the TF-specific free parameters (β_k, α_k and ω_k) by simultaneous fits of the model to the 37 *D. melanogaster* CRMs. To that effect, we use the Nelder-Mead simplex model for local optimization and random restarts to reduce the effect of local optima. This is the same procedure used by He et al. (2010). Figure 3.1 shows the real patterns driven by a set of AP CRMs, as well as the pattern predicted using GEMSTAT. As it can be noticed, the fits from GEMSTAT capture the expression domains of most CRMs without overfitting the data.

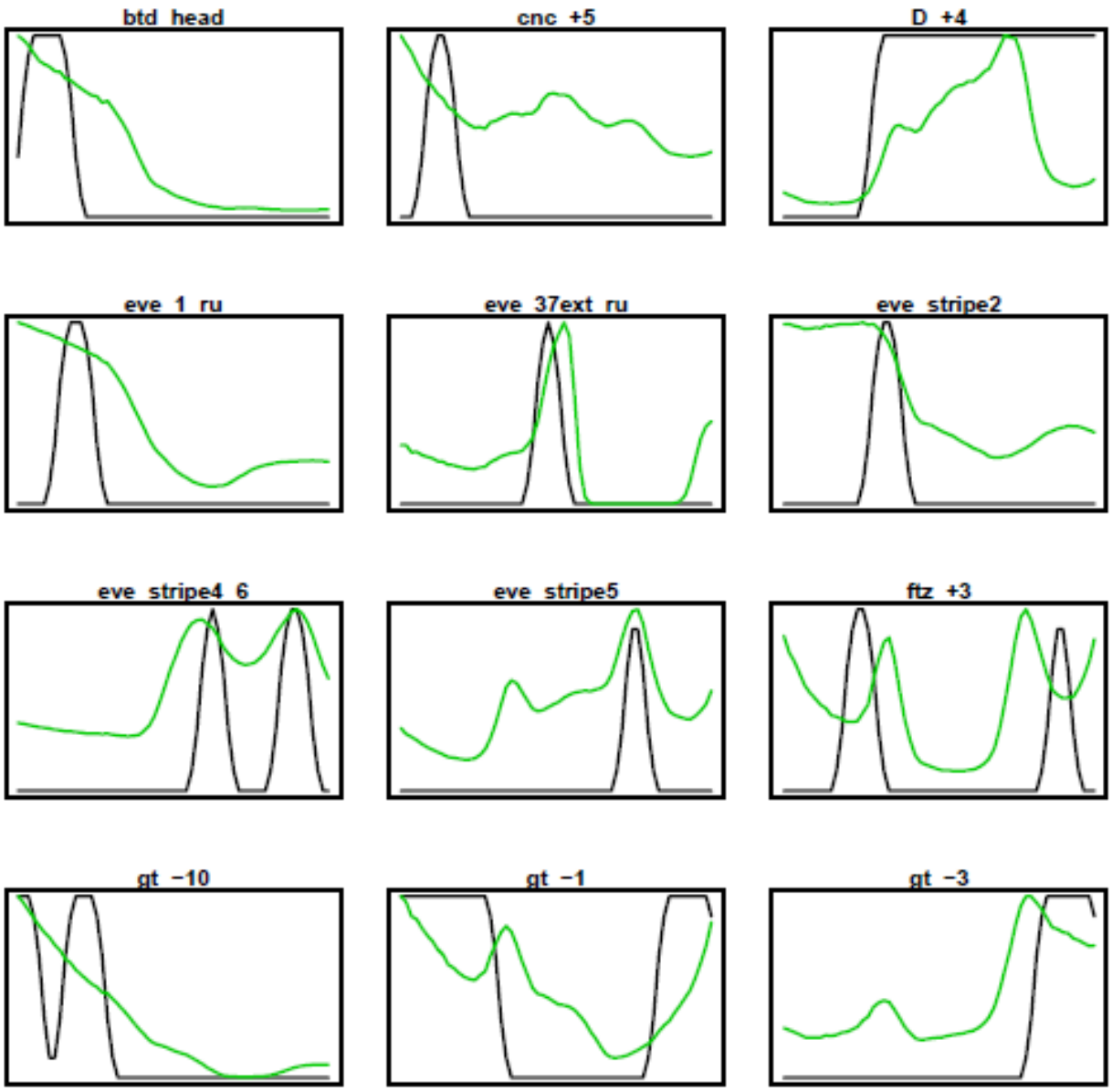


Figure 3.1: Example of GEMSTAT Fits. Expression pattern predicted by GEMSTAT is shown in green, for 15 CRMs whose real patterns are shown in black.

By default, we configured GEMSTAT to use self-cooperativity for only for BCD and KNI, as this model had been found to be the optimal model by He et al. (2010). This means that for all other TFs the self-cooperativity parameter was set to 1 (no effect on statistical weight). Additionally, we limit the values β_k, α_k to a range of biophysically plausible values and, in particular, limit α_k

to values that are appropriate for the known role of the TF (i.e. $\alpha_k < 1$ for repressors and $\alpha_k > 1$ for activators).

We perform the GEMSTAT model fitting step independently of the evolutionary simulations and use the same 37 *D. melanogaster* CRMs for our evolutionary simulations described in Chapters 4, 5, 6 and 7. Therefore, the parameter fitting for GEMSTAT is not influenced by the results of the evolutionary simulation. For some of the experiments, we performed simulations with a different configuration of GEMSTAT (e.g., self-cooperativity for a different subset of TFs, model including a different set of TFs). In these cases, the free parameters of the model are re-trained on the same data set, again independently of the fits to the evolutionary simulations.

3.2 The fitness function: wPGP

The next component of the PEBCRES model is the fitness function, which compares the predicted expression pattern for a CRM sequence with the ideal expression pattern. A desirable fitness function has three properties: (1) the value of the fitness function is maximal when the predicted expression pattern is a perfect match to the ideal expression pattern, (2) any deviation from the ideal expression pattern is penalized by decreasing the fitness function, and (3) the penalty value is monotonically increasing with the amount of deviation.

With these considerations in mind, we used the "weighted Pattern Generating Potential" (wPGP) score of Samee and Sinha (2013) to compare the ideal expression with a predicted expression. Let u be the pre-determined ideal expression profile for the CRM. Let v represent the expression profile predicted by GEMSTAT for a given genotype g, and let u_i and v_i represent the ideal expression and predicted expression respectively in cell type i. The fitness of genotype g is defined from the wPGP score between u and v, as follows:

1. Compute $reward$ as $$\frac{\sum u_i \times \min(u_i, v_i)}{\sum u_i^2}$$

2. Compute $penalty$ as $$\frac{\sum (u_{max} - u_i) \times \max(0, v_i - u_i)}{\sum (u_{max} - u_i)^2}$$

3. Compute $wPGP$ as $$wPGP(u) = reward - penalty$$

4. Compute fitness functional as $$f(g) = [\max(0, wPGP(u))]^2$$

5. Compute *fitness* as $$F(g) = 1 + Kf(g)$$

where $u_{max} = \mathrm{Max}_i[u_i]$ and K is a free parameter representing the selection scaling constant. Note that $wPGP$, and therefore $F(g)$, is maximized when the reward is maximum and the penalty is minimum. The reward is maximized when $v_i \geq u_i$ for every cell-type i; in other words, a sequence is rewarded for driving higher expression in cell-type i. On the other hand, penalty is minimized when $v_i \leq u_i$, or, in other words, over-expression in any cell-type i is penalized. Putting reward and penalty together, we have that fitness $F(g)$ is maximized when $v_i \geq u_i$ and $v_i \leq u_i$, which only happens when $v_i = u_i$. Having $v_i > u_i$ for any cell-type i increases penalty while reward remains the same and having $v_i < u_i$ decreases reward with penalty remaining the same. The wPGP score, and thus also the fitness function has all the aforementioned properties. Additionally, over-expression and under-expression are penalized differently, and over-expression is penalized only up to a saturation point. The advantages of the wPGP score over either the sum of squared errors or a correlation coefficient are discussed in Samee and Sinha (2013). The *fitness functional* $f(g)$ is a number between 0 and 1, with a value of 1 representing perfect match between u and v. The parameter *K* can be interpreted as the selection coefficient when the two competing genotypes have $f(g)$ equal 0 and 1 respectively.

Notice that the proposed fitness function addresses all of the common issues listed on section 1.5.1 above. First of all, it defines a continuous fitness landscape (issue 1), where changes to a

binding site that result in a small change in binding affinity will cause small changes to the predicted expression pattern and therefore the relative change in fitness will be small. Moreover, the fitness function weights the individual contributions of all of the binding sites in the sequence, estimating the relative importance of each individual binding site while accounting for the expression profile driven by the CRM, therefore addressing issue 2.

Issue 3 relates to changes in the context of a binding site; and our fitness function addresses it in two ways: first, changes to any functional binding site in the sequence will cause a change in the predicted expression pattern and therefore, change in the fitness function. Second, changes in the background that create a new site will also affect the fitness function since GEMSTAT uses a very low threshold on binding site calling, as discussed in section 3.1.1. Finally our fitness function is capable of dealing with the combinatorial nature of gene regulation since GEMSTAT models the diverse roles and expression profiles or different TFs, abstracting the combinatorial logic in the expression pattern prediction.

3.3 The simulation model: Wright-Fisher

The Wright-Fisher model (Wright 1931; Hartl and Clark 1997; Fisher 1999; Hein et al. 2004) is a simple model of the evolution of a population. It describes the evolution of a simplified population as individuals transition from generation to generation. In its simplest form, the Wright-Fisher model evolves a population of *N* diploid individuals by simulating the evolution of 2*N* alleles representing haploid individuals. (There are variations of the model that explicitly represent diploid populations; however, we have opted for this simpler and more common version).

Figure 3.2 is cartoon of the Wright-Fisher model, as used in PEBCRES. In our model a population of 2*N* individuals (yellow circle in Figure 3.2), each represented the sequence of a putative CRM, evolves through a series of discrete generations t_1, t_2, …, t_n. The population is initialized with 2*N* copies of the same initial sequence (in generation t_1) and every subsequent generation is sampled, with replacement, from the previous generation. An individual in generation t_i may

spawn any number of copies in generation t_{i+1} (black arrows in Figure 3.2), with the only restriction being that the population at every generation be of size 2*N*.

The expected number of copies an individual spawns in the next generation is determined by the fitness of that individual relative to the population. Recalling that the fitness of an individual i is $F(i) = 1 + Kf(i)$, and individual with *fitness functional* $f(i) = 1$ is expected to generate (1+*K*) times the number of copies of an individual j whose *fitness functional* is $f(j)$ equal to 0. In general an individual i is expected to generate a number of copies proportional to $F(i)$.

At any generation, after a new population has been sampled from the previous population, a mutation process is simulated, possibly changing the sequence of one or more individuals. Any individual modified by a mutation needs to have its fitness recomputed by predicting its expression profile using GEMSTAT and calculating $F(\cdot)$ according to the procedure outlined on section 3.2. In our model, mutations are usually limited to single nucleotide substitutions, however, insertions, deletions and tandem repeats are also implemented and we have used these in experiments published in He et al. (2012) and Duque et al. (2013).

Notice that the Wright-Fisher model, at least as used in PEBCRES, makes a number of simplifying assumptions, the most important of which are the existence of discrete, non-overlapping generations, with fixed population size, no geographical or social structure and the complete absence of recombination. In practice, however, these assumptions do hold. For example, assuming discrete, non-overlapping generations would be equivalent to assuming that all individuals reproduce at the exact same time and immediately die, and that the number of offspring is exactly the same as the number of individuals in the generation before. Nevertheless, theories based on the Wright-Fisher model (e.g. the coalescence theory (Hein et al. 2004)) are robust to these assumptions and the simplicity of the model allows for efficient simulation and easy mathematical manipulation.

3.4 The PEBCRES model

The PEBCRES model is illustrated in Figure 3.2 and summarized in Algorithm 3.1.

Algorithm 3.1: The PEBCRES model

01 Start with a population of 2N identical individuals at simulation time $t_{now} = 1$
02 Sample next mutation time t_{mut}
03 Repeat:
04 If the simulation time $t_{now} = t_{mut}$:
05 Choose a random individual i and mutate it
06 Evaluate the fitness $F(i)$ of the mutate individual
07 Sample a new mutation time t_{mut}
08 End if
09 Sample a new population with 2N individuals
10 Update the simulation time $t_{now} = t_{now} + 1$
11 *Until* STOPING CRITERIA *is met*

As it can be seen in Algorithm 3.1 the structure of the PEBCRES model is dictated by the Wright-Fisher model described in Section 3.3. The essential difference between PEBCRES and other evolutionary simulators based on the Wright-Fisher model is step 06. Often times it is difficult to estimate what the effect of a mutation is on the fitness of the individual, but PEBCRES uses the fitness function defined in section 3.2, which addresses the common issues of models of regulatory sequence evolution (see section 1.5.1).

Mutations (step 05) are typically point substitutions but can also include insertions, deletions and tandem repeats. If the mutation is a point substitution than a random position in the sequence is selected and the nucleotide in that position is changed to one of the three remaining nucleotides, with uniform probability. If insertions and deletions are enabled, one must first choose whether the mutation is a point substitution or an indel. Indels have a probability of 20%, and if the mutation is chosen to be an indel the model chooses the type of indel with a 60% deletion bias. Both insertions and deletions have length distributions sampled from a mixture of two geometric distributions, with parameters learned from data by Jaebum Kim (personal communication). Insertions can be either a repeat from the same CRM or a

randomly generated sequence, with the mode of insertions chosen before the simulation as a global parameter.

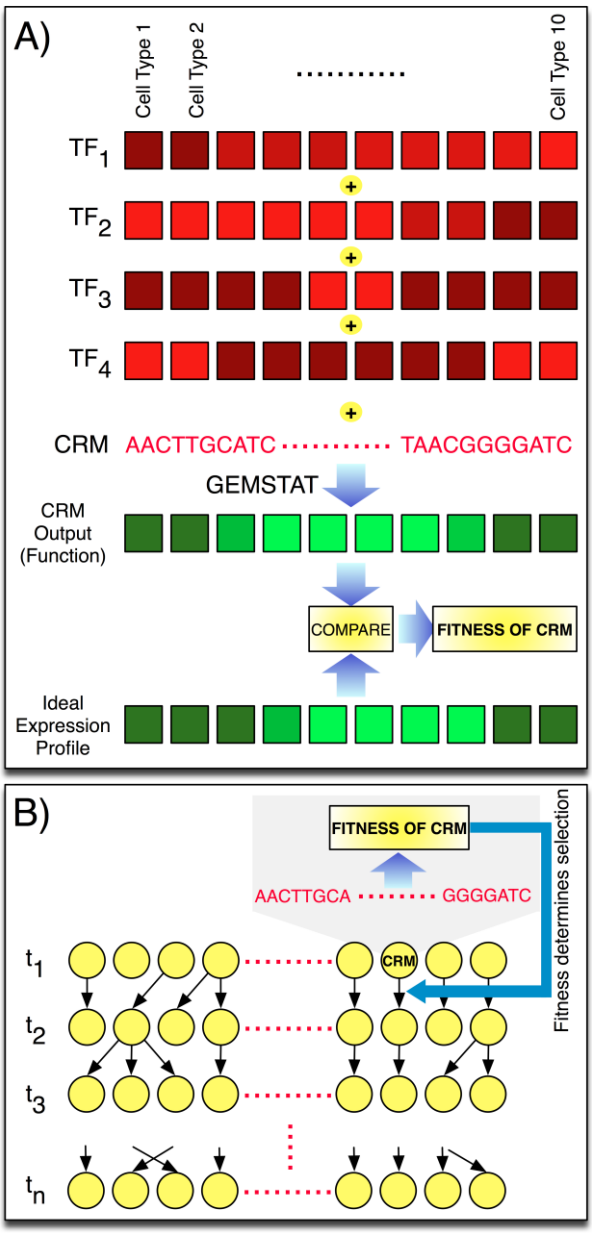


Figure 3.2: Illustration of PEBCRES. (A) The regulatory function of a CRM is represented by an "expression profile", that is, gene expression levels in well-defined cell types. An ideal expression profile (shown in green, with brighter shades representing higher expression) is designated, and the fitness of a CRM sequence is computed by comparing this ideal expression profile to that predicted as being the CRM's output. (A more similar CRM output profile has greater fitness.) The CRM's output is computed based on its sequence and the concentration values of relevant TFs (shown in red) in the same set of cell types. This computation is done using the thermodynamics-based GEMSTAT model (He et al. 2010), which additionally uses the binding motifs of those TFs to predict CRM function from sequence. **(B)** Cartoon illustration of Wright-Fisher simulations underlying the PEBCRES model. A fixed-sized population of individuals (CRMs) is evolved for *n* generations (t_1, t_2, … t_n). Random mutations are introduced in each generation using a pre-determined mutation rate parameter. Each individual is sampled independently at random from the population in the previous generation, and this sampling probability is dependent on the fitness of the individual, which in turn is determined by the CRM's output as shown in **(A)**.

The stopping criteria (step 11) is typically a limit on the number of generations or check for evolutionary divergence between the current population and the founding individual, but other criteria are also possible, such as fitness threshold or number of generations without significant change in fitness.

3.4.1 Inputs and parameters

This section briefly describes the parameters and inputs required by PEBCRES. First of all PEBCRES requires a complete GEMSTAT specification, defined henceforth as the set of all of the information needed to predict the expression driven by a sequence. The GEMSTAT specification includes the binding specificity (PWM) and expression profiles of all of the relevant TFs, as well as the value of all of the free parameters of the GEMSTAT model (α_k, β_k and ω_k). The complete GEMSTAT specification is usually obtained by first choosing the set of TFs and cooperativity assumptions and, subsequently, training the free parameters of GEMSTAT as described in section 3.1.2. Using different GEMSTAT specifications allows PEBCRES to explore the effects of various assumptions regarding the mechanisms of gene regulation.

Next, PEBCRES requires an initial sequence, which is used as the starting point for the evolutionary simulation (with all individuals set to the same sequence), and a target expression pattern, which is used in the fitness function. This pair of inputs is used to specify whether the mode of evolution is adaptive or purifying selection. To simulate evolution under purifying selection we provide PEBCRES with an initial sequence that drives the target expression pattern, resulting in an initial fitness equal to 1. On the other hand, for adaptive selection we typically provide PEBCRES with a random initial sequence resulting in a low initial fitness (close to 0), which evolution will attempt to improve.

Finally PEBCRES has a set of population genetics parameters, namely the population size (*N*), the selection scaling constant (*K*) (see section 3.2) and the mutation rate (μ). The selection scaling constant (*K*) is of special relevance as it is the hardest parameter to estimate experimentally, and is instead fit to data on Chapter 4.

3.4.2 Fast simulation with time rescaling

In our simulations using PEBCRES we use time rescaling (Hoggart et al. 2007) to speed up the simulation and reduce use of computational resources. Using time rescaling we can simulate a population of size $2N_r$ evolving for T_r generations with mutation rate μ_r and selection coefficient s_r by evolving a population of size $2N_s = \frac{2N_r}{\lambda}$ for $T_s = \frac{T_r}{\lambda}$ generations with mutation rate $\mu_s = \mu_r \times \lambda$ and selection coefficient $s_s = S_r \times \lambda$, where λ is the time rescaling parameter. Note that the subscript (*r*) indicates the real parameter and the subscript (*s*) indicates the scaled parameter.

The advantage of using time rescaling is that it allows us to simulate real populations for real time scales while using a significantly smaller population size and time and while maintaining the important population genetics quantities ($4Ns$ and $2N\mu$) unaltered. The smaller population size and number of generations imply that the computational time necessary to simulate the evolution is significantly shorter and that memory requirements are also reduced.

Note that standard values for Drosophila populations in the literature are in the range of $10^5 - 10^6$ for population size ($2N$)(Thornton and Andolfatto 2006) and $10^{-9} - 10^{-8}$ for mutation rate (μ)(Drake et al. 1998), resulting in $2N\mu$ in the range of $10^{-2} - 10^{-4}$. Therefore we can simulate the evolution of Drosophila CRMs using, for example, parameters $2N = 100$ and $\mu = 10^{-5}$, corresponding to a time rescaling parameter $\lambda = 1000$ and resulting in $2N\mu = 10^{-3}$.

Beyond the time rescaling, PEBCRES also includes coding optimizations to reduce memory footprint and execution time. For example PEBCRES does not explicitly store the complete sequence of every individual, but rather, stores the evolutionary history of the individual and reconstructs the sequence from this information. The code for PEBCRES, which is written in C++ for efficiency reasons, is available at: (http://veda.cs.uiuc.edu/evolsimul/).

3.5 Summary and conclusions

In this chapter we introduced PEBCRES, a realistic, flexible and efficient model for the simulation of the evolution of regulatory sequences. PEBCRES is composed of three components: a sequence-to-expression model (section 3.1), a fitness function that uses the sequence-to-expression model to evaluate putative CRMs (section 3.2) and an evolutionary model that simulates mutations and their evolutionary fates (section 3.3). We have also described how these components are combined to form PEBCRES (section 3.4) and what are the necessary inputs and parameters for our model (section 3.4.1).

We describe PEBCRES throughout this thesis as realistic, efficient and flexible for the following reasons: First, it is efficient since it uses an efficient sequence-to-expression model as well as time rescaling and several implementation optimizations. Second, it is flexible since it is capable of simulating diverse modes of evolution (e.g. adaptive, negative selection) and a variety of mechanistic assumptions (e.g. different cooperativity assumptions). It is also capable of evolving a variety of expression patterns. And finally, it is realistic because it assumes that the fitness of a sequence is dependent on the expression pattern driven by that sequence, and not based on simplified models (e.g. maximizing total activation output (MacArthur and Brookfield 2004)).

In the following chapters of this thesis we will endeavor to demonstrate that PEBCRES is more accurate than existing models of regulatory sequence evolution as well as to showcase how PEBCRES can be used to acquire new and relevant insights on evolution and gene regulation.

4 Modeling evolutionary data

In this chapter we describe the use of PEBCRES and other tools to address our first biological objective: accurately modeling evolutionary data. The objectives of this chapter are two-fold: first we demonstrate the accuracy of our simulation software; second we demonstrate the importance of addressing the common issues of models of regulatory sequence evolution described in section 1.5.1 . The chapter is organized as follows: section 4.1 formalizes the problem we are trying to address and discusses its relevance, as well as other approaches from the literature. Next, section 4.2 presents the results of applying PEBCRES to the task of modeling evolutionary data. Finally section 0 summarizes and concludes this chapter.

4.1 Background and motivation

This thesis proposes the use of a computational simulation of regulatory sequence evolution to address questions related to the evolution and mechanisms of gene regulation. However, before we can test hypotheses regarding mechanistic features of gene regulation or draw conclusions regarding evolutionary artifacts or the time necessary to evolve a CRM, it is first necessary to demonstrate that the simulations performed using PEBCRES are actually accurate. In this chapter we will demonstrate this by modeling the evolution of 37 CRMs from *D. Melanogaster*. These CRMs are responsible for driving expression in the AP system in Drosophila, and are therefore under strong selection for conservation. Kim et al. (2009) studied the evolution of these and other developmental CRMs by comparing theirs sequences across 12 Drosophila species.

Kim et al. (2009) used a number of different statistics to compare the conservation patterns across orthologous sequences for different TFs or groups of binding sites. Their main objective was different from the objective of this chapter; they mainly sought to find out how different features of binding sites or CRMS influence the conservation patterns, while we seek to validate our computational simulation tool. However, in most of their exercises, Kim et al. (2009)

[*] This chapter contains material previously published in (Duque et al. 2013).

presented summary statistics that represent the real pattern of evolution of regulatory sequences in the Drosophila species. In particular, Kim et al. (2009) described the expected level of conservation for binding sites of seven TFs: *Bicoid* (BCD), *Caudal* (CAD), dSTAT, *Tailless* (TLL), *Hunchback* (HB), *Kruppel* (KR), and *Knirps* (KNI). In this chapter we will use a similar methodology to demonstrate that our software models the evolution of regulatory sequences in an accurate manner.

Additionally, Kim et al. (2009) tried to model these patterns of conservation using two models: The SS model (Kim et al. 2009) and the HB model (Halpern and Bruno 1998; Kim et al. 2009) to precisely that type of data. We will use these models as a point of comparison to demonstrate that or tools represent an improvement over the previous state-of-the-art evolutionary models. In this chapter we also introduce Predicted-Expression-Based Site Evolution Simulator (PEBSES), a simpler evolutionary model designed to closely match the SS model and to demonstrate the importance of different aspects of the PEBCRES model.

4.2 Results

In this section we will demonstrate that PEBCRES model is capable of reproducing and/or predicting the conservation patterns in regulatory sequences with better accuracy[2] than existing models of regulatory sequence evolution. We will also demonstrate the importance of addressing the common issues of evolutionary models of regulatory sequence evolution introduced in section 1.5.1 by first demonstrating the improvements in accuracy that can be gained by addressing issues 1 (*Continuous nature of the functional effect of mutations*) and 2 (*Context in which a binding site evolves*) only. Subsequently, we will demonstrate how further improvements can be gained by additionally addressing issues 3 (*Evolutionary changes in the context of the binding site*) and 4 (*Combinatorial Regulation by Multiple TFs*) in PEBCRES.

[2] The meaning of accuracy in this context will be formalized in section 4.2.2.

4.2.1 The PEBSES model

In this section we introduce PEBSES, a simplified model based on the PEBCRES model described in Chapter 3 and on the SS model from Kim et al. (2009). The PEBSES model is designed to address common issues 1 (*Continuous effect of mutations*) and 2 (*Binding site context*, see section 1.5.1) and at the same time not address issues 3 and 4 (*Evolving context* and *Combinatorial regulation*)[3]. In other words, PEBSES is designed to demonstrate the relevance of our fitness function introduced in section 3.2, in isolation from other components of PEBCRES.

Recall that the fitness function introduced in section 3.2 was specifically designed to be sensitive to arbitrary changes in gene expression caused by changes to the regulatory sequence. The change in transcription output caused by a mutation, however small this change may be, is captured by our model in a continuous manner; with less relevant changes (e.g. changes to a non-specific position of a binding site) having a smaller impact on fitness. This characteristic of our fitness function addresses issue 1 (binary fitness function). Moreover, the fitness of a sequence is dependent on the expression driven by that sequence with every putative binging site, strong or weak, influencing that pattern. This characteristic addresses issue 2 (sensitivity to context). Finally our fitness function estimates the transcriptional effect of changes to any part of the sequence, including all putative binding sites as well as background sequences, i.e., sequences with too little affinity to the PWM of any TF to be considered "functional". This is an important consideration since background sequences can accumulate mutations and become functional. Our final consideration addresses issue 3 (changes in the context).

The PEBSES model uses our fitness function in conjunction with a continuous-time Markov process to simulate the fixation of substitutions on a single binding site following the theory of Kimura and Otha (Kimura and Ohta 1969; Mustonen and Lässig 2005; Kim et al. 2009). It simulates the evolution of a single binding site while considering the context in which that

[3] See section 0 for a discussion on the relevance of this particular design objective.

binding site is located by using our fitness function. As a continuous-time Markov chain simulator, PEBSES models evolution using a single sequence representing the consensus population (as opposed to explicitly representing a population of individuals like PEBCRES). At every step of the simulation, a collection of putative sites (b) is generated by enumerating all possible one nucleotide substitutions to the current binding site (a). One of the putative sites (b) is chosen with probability proportional to the rate $u(a, b)$.

The rate of substitution $u(a, b)$, from a site a to a site b is given by the following equation:

$$u(a,b) = 2N\mu(a,b)\frac{1 - e^{[-2(\mathcal{F}(b)-\mathcal{F}(a))]}}{1 - e^{[-4N(\mathcal{F}(b)-\mathcal{F}(a))]}}$$

where N represents the effective population size, $\mu(a, b)$ is the background rate of mutation from site a to site b (Kimura 1980) and $\mathcal{F}(x)$ is the fitness of a site x relative to the current site, i.e., $\mathcal{F}(x) = F(x)/F(a)$, with $\mathcal{F}(a) = 1$ by definition. This leads to $\mathcal{F}(b) - \mathcal{F}(a) = \frac{K(f(b)-f(a))}{1+Kf(a)}$, which we can approximate as $\mathcal{F}(b) - \mathcal{F}(a) \approx K(f(b) - f(a))$, when $Kf(a) \ll 1$. This will be the case in our simulations since $f(a) \leq 1$ and $K \ll 1$ (see section 3.2 for a distinction between $F(x)$ and $f(x)$).

After a substitution has been selected, the current binding site is updated to reflect the mutation and the simulation time is incremented by a random number sampled from an exponential distribution with rate $U = \sum_b u(a, b)$, where the summation is over every site b in the collection of all putative sites that differ from a by exactly one nucleotide. This procedure is equivalent to assuming that each of substitutions is a competing process with expected time exponentially distributed with rate $u(a, b)$ and that we are waiting for the first of these processes to happen. This is a typical procedure in population genetics simulations, like the coalescence processes (Hein et al. 2004).

The Markov process ends when the simulation time is greater than or equal to some predetermined value, which in our experiments was set to match the evolutionary distance between *D. melanogaster* and *D. yakuba*.

Comparison with the SS model: PEBSES is, by design, analogous to the SS model (Kim et al. 2009) in all aspects except for the fitness function $\mathcal{F}(x)$. Indeed, the SS model can be implemented inside PEBSES simply by redefining $\mathcal{F}(x)$ as:

$$\mathcal{F}(x) = \begin{cases} 1+s & \text{iff LLR}(x) > \theta \\ 1 & \text{otherwise} \end{cases}$$

where LLR(x) represents the log-likelihood ratio of binding site x determined by the match to the TF PWM, θ is the threshold on site strength used by the SS model and $s > 0$ is the selection coefficient. However, contrary to the SS model, PEBSES addresses common issues 1 (*continuity of fitness effects*) and 2 (*sensitivity to binding site context*) from section 1.5.1.

Limitations of the continuous-time Markov process: Despite the fact that our fitness function is capable of addressing all of the common issues from section 1.5.1, the PEBSES model does not address common issue 3 (*sensitivity to changes in context*), since PEBSES is a binding site evolution model, with changes restricted to a single binding site. PEBSES also does not address issue 4 (*Combinatorial nature of gene regulation*), once again due to the restriction to simulating the evolution of a single binding site.

These restrictions are both due to limitations inherent from continuous-time Markov process used in PEBSES, which limits the simulation to sequences of relatively short length. The continuous-time Markov process used in PEBSES (as well as the SS model) simulates successive fixation events and assumes that during the time necessary for a mutation to fix in the population, no other competing mutation arises. This assumption is only valid if the fixation time is smaller the waiting time for new mutations, which is true if the mutation rate is small and selection is strong. However, the mutation rate scales with the length of the sequence and therefore the key assumption in the Markov process used in this section is only valid for short sequences.

Additionally, the simulation of the continuous-time Markov process requires evaluating the fitness of every single point mutation to the sequence, a process that is computationally inefficient for long sequences.

Relevance of the PEBSES model: PEBSES represents an intermediary step between the models of Lässig and co-workers (Berg et al. 2004b; Mustonen and Lässig 2005) and Kim et al. (2009) and the PEBCRES model introduced in this thesis. The PEBCRES model was designed to address four common issues of models of regulatory sequence evolution and to be as realistic and flexible as possible; however, it is not applicable to every conceivable question a biology researcher might be interested in. This is made evident by the variety of forms that the models of the evolution of gene regulation have taken (e.g. (Stone and Wray 2001; Carter and Wagner 2002; Berg et al. 2004b; Mustonen and Lässig 2005; Durrett and Schmidt 2007; Francois et al. 2007; Durrett and Schmidt 2008; Cooper et al. 2009; Kim et al. 2009; Josephides and Moses 2011; Stewart et al. 2012), just to cite a few).

Therefore, even though PEBCRES might be applicable to many open questions from regulatory genomics (e.g. Chapters 5, 6 and 7), new models will still need to be designed for a variety of other questions. Whoever designs such models will have to carefully consider which issues to address and which issues to ignore and the PEBSES allows us to understand which improvements can be achieved and which insights can be learned by addressing issues 1 and 2 only, while the PEBCRES model demonstrates the usefulness of addressing issues 3 and 4. Additionally the PEBSES is more efficient in terms of computational resources than the PEBCRES model when simulating the evolution of a single binding site is better equipped to implement insights from continuous-time population genetics theories like the coalescent theory (Hein et al. 2004). .

4.2.2 Evaluating PEBSES on evolutionary data

In this section we describe the experiments designed to evaluate how effective our approach is at modeling evolutionary data. Each of our experiments consists of several simulations of the evolution of one out of 37 CRMs from the AP system, under purifying selection.

In each simulation one binding site from *D. melanogaster* is selected at random among the set of all high confidence binding sites in our collection of *D. melanogaster* CRMs. The set of high confidence binding site is obtained using the procedure from Kim et al. (2009). The target expression pattern throughout the simulation is the pattern predicted by GEMSTAT for that CRM (and hence evolution proceeds under purifying selection) and the parameters for GEMSTAT have been trained to exhibit strong agreement between the prediction and the experimentally determined expression pattern (He et al. 2010). Note that our choice of target pattern throughout the simulation reflects the assumption each CRMs in *D. melanogaster* and *D. yakuba* drives the same expression pattern. This assumption is reasonable for simulations of CRMs from the AP system (Hare et al. 2008; Weirauch and Hughes 2010; Swanson et al. 2011).

We adopted the methodology of Kim et al. (2009) to generate summary statistics of binding site conservation for each of five different TFs – *Bicoid* (BCD), CAD, *Hunchback* (HB), *Kruppel* (KR), and *Knirps* (KNI) – in 37 different cis-regulatory modules from *D. melanogaster*. These CRMs were selected because each of them is associated with an experimentally characterized expression profile, and because these experimental profiles can be predicted with moderate accuracy from respective sequences by the GEMSTAT model (Supplementary Fig. S1 of (He et al. 2012)). To generate descriptive statistics of binding site conservation, (1) *D. melanogaster* CRM sequences were aligned to orthologous *D. yakuba* sequences, (2) for every predicted binding site in a *D. melanogaster* sequence, the binding energy of the site and its orthologous site was predicted using the TF's motif (PWM), (3) the difference in computed binding energies was noted as the "energy difference", and (4) a histogram of energy differences of orthologous sites was created by examining sites across all CRMs. (See Supplementary Methods for brief overview.) This histogram serves as the evolutionary data to be modeled. An analogous histogram was computed based on the simulations of evolutionary models such as the Halpern-Bruno model, the SS model, or PEBSES model, and compared with the evolutionary data. The results (Figure 4.1) show that PEBSES models the evolutionary data more accurately than either the SS or the Halpern-Bruno model. For instance, when focusing on sites of the TF KR (Figure 4.1 (C)), the energy difference histogram from PEBSES predictions is in strong agreement with

that from real data, as measured by the Kolmogorov-Smirnoff (KS) d-statistic. The fits are significantly worse for the SS and Halpern-Bruno models. The same is true for sites of the TF CAD (Figure 4.1 (B)). For BCD sites, the PEBSES and SS models have comparable values of the KS d-statistic (Figure 4.1 (A)). Results for HB and KNI sites are not shown, but in both cases the PEBSES model has the lowest KS d-statistic value, implying better fits. An alternative, more compact way to compare the different models is to plot the fraction of sites for which energy difference is 0 (indicating perfectly conserved sites), in the real data as well as in simulations under each model. Figure 4.1 (D) shows that by this criterion the PEBSES model makes the most accurate predictions of the observed evolutionary characteristics of binding sites. Its overall error is the lowest of the three models compared and it makes the most accurate predictions for CAD, KR, and KNI sites. (For BCD and HB sites, the best fits belong to the SS and Halpern-Bruno models respectively.)

We note that PEBSES uses only one free parameter (see Methods), while the SS model uses one free parameter per TF (Kim et al. 2009). In addition to being a more constrained model, PEBSES is arguably a more realistic model of binding site evolution. It uses a state-of-the-art sequence-to-expression model to assign fitness to sequences, and this underlying model is in turn trained on sequence and expression data for a large number of CRMs. Moreover, it is easy to change the underlying model used in PEBSES and simulate the evolution of a site under different mechanistic assumptions, e.g., cooperativity between TFs, short range repression, synergy activation, etc. We explore this feature in a later section. The sequence-to-expression model used in these simulations incorporates self-cooperative DNA-binding by BCD and KNI (as suggested in He et al. (2010)), but no other TF.

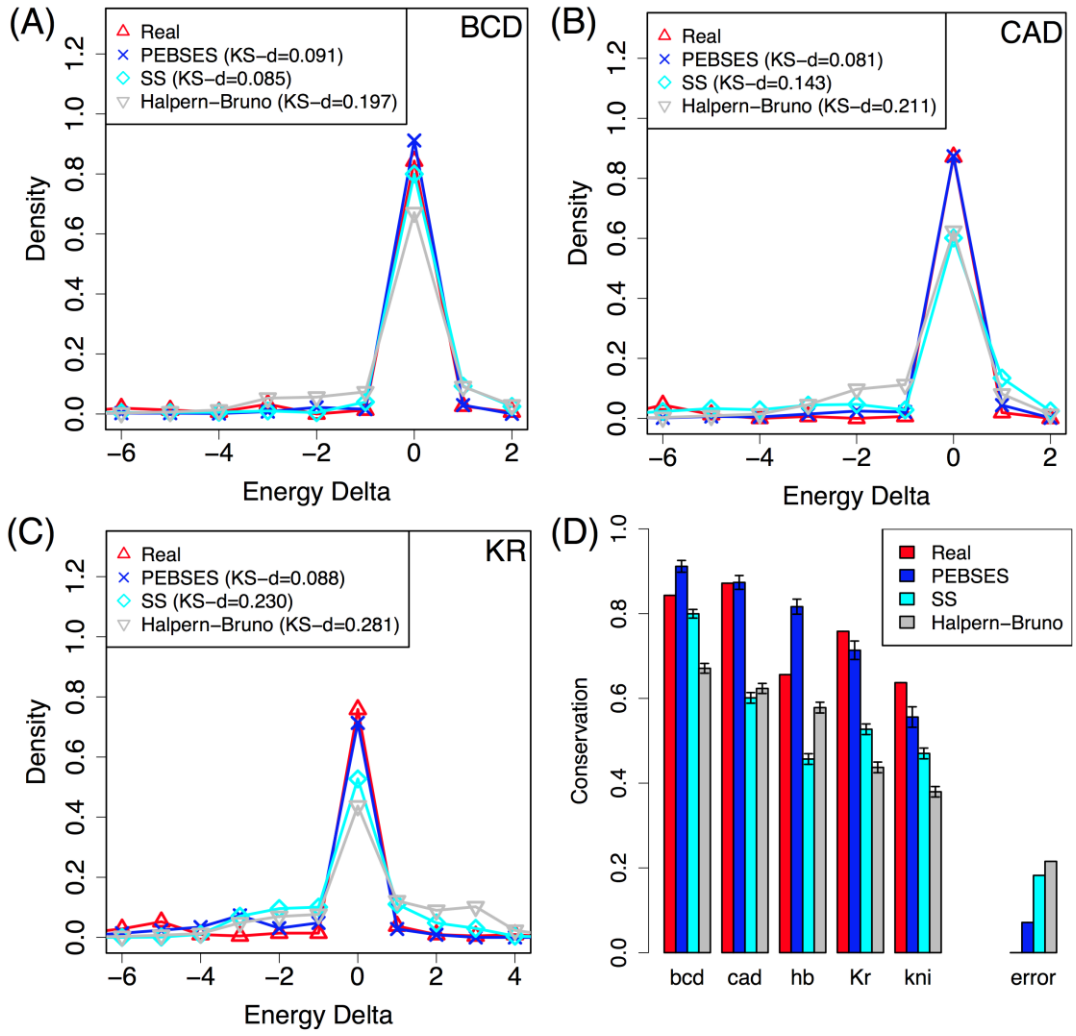


Figure 4.1: Comparison between PEBSES and existing models.

(A-C) Energy difference histograms from real data and from three evolutionary models – Halpern-Bruno (Halpern and Bruno 1998), SS (Kim et al. 2009), and PEBSES (section 4.2.1) – for binding sites of TFs BCD **(A)**, CAD **(B)** and KR **(C)**. A binding site in a *D. melanogaster* CRM was compared to its aligned site in *D. yakuba* (for real data histogram), or in a simulated descendent (for model-based histograms), and the difference in predicted binding energies (LLR scores) of the two sites was noted. This was repeated for each of 159, 171, and 239 binding sites of BCD, CAD and KR. For model-based histograms, each site's evolution was simulated on an average of 28 times. **(D)** The fraction of sites for which energy difference between *D. melanogaster* and *D. yakuba* orthologs is zero ("Conservation", y-axis) is shown for real data and for the Halpern-Bruno, SS and PEBSES models, and for five different TFs. The difference between real data and a model's prediction of this fraction is deemed the TF-specific error of that model, and the absolute value of error is averaged over the five TFs and shown as the "error" of each model.

The results above suggest that the context of a site, represented by the sequence surrounding it and the expression driven by the sequence, plays an important role in the evolutionary dynamics of that site. Capturing this role in simulating binding site evolution leads to better fits to evolutionary data. It has been speculated that phenomena such as homotypic clustering may buffer a CRM against mutation in the sites (Spivakov et al. 2012), thereby increasing the frequency of in-site mutations. Our findings above reveal a complementary phenomenon, that is, that the context of a site can also increase the selection pressure for conservation at the site.

4.2.3 Evaluating PEBCRES on evolutionary data

In the previous section demonstrated that PEBSES is capable of modeling conservation patterns between *D. melanogaster* and *D. yakuba* with better accuracy than the SS model and the HB model, while at the same time requiring the estimation of fewer free parameters. However, as discussed in section 4.2.1, PEBSES does not address issues 3 (*sensitivity to changes in context*) and 4 (*combinatorial regulation*). Moreover, PEBSES is not a population based model, and is therefore unable to model several relevant population genetics phenomena such as competing polymorphisms (Barreiro et al. 2008) and structural variations (Feuk et al. 2006) or genetic draft (Neher 2013). In this section we evaluate PEBCRES, the simulation software presented in this thesis, on the same task described in section 4.2.2.

PEBCRES addresses PEBSES's drawbacks by explicitly representing the population following the Wright-Fisher (Hedrick 2011) model, with fixed population size. By explicitly representing the population instead simulating the continuous-time Markov process, PEBCRES can simulate the evolution of arbitrarily long sequences like a CRM, instead of being restricted to short sequences like a binding site. PEBCRES is described in details in Chapter 3.

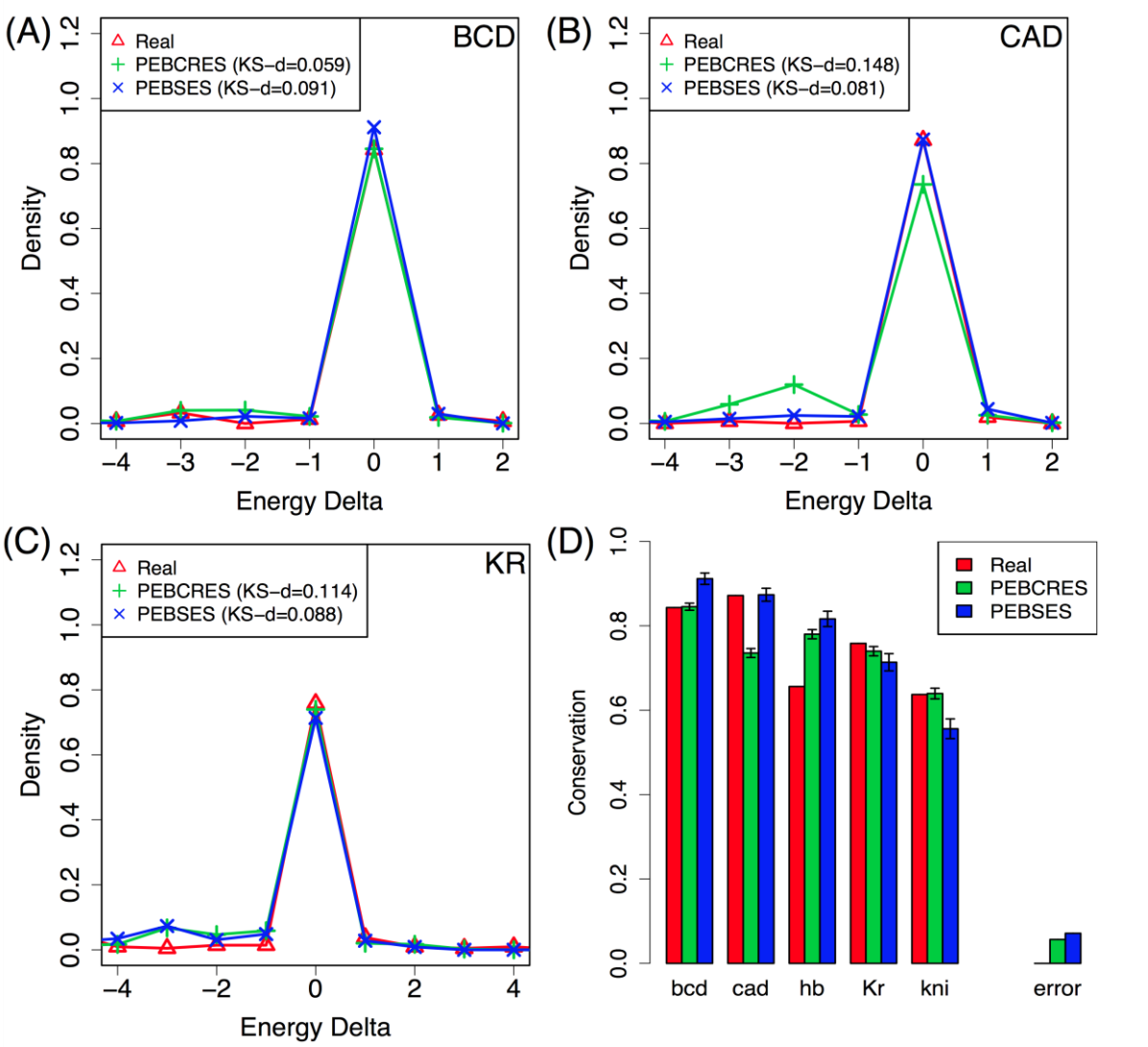


Figure 4.2: Comparison between PEBSES and PEBCRES. (A-C) Energy difference histograms from real data and from the two evolutionary models – PEBSES and PEBCRES – presented in this work, for TFs BCD **(A)**, CAD **(B)** and KR **(C)**. **(D)** The fraction of sites for which energy difference between *D. melanogaster* and *D. yakuba* orthologs is zero ("Conservation", y-axis), shown for real data and for the PEBSES and PEBCRES models. The error of either model, as defined in legend of Figure 4.1, is also shown.

Figure 4.2 compares the histogram of binding site energy differences from PEBCRES and PEBSES simulations to evolutionary data. We find that PEBCRES simulations provide significantly better fits to data on BCD and KNI sites, and significantly worse fits for CAD sites, while both models exhibit similar levels of agreement with data on HB and KR sites. We also performed a set of

PEBCRES simulations that included insertions and deletions ("indels") as evolutionary events using indel rates and length distributions suggested in the literature (He et al. 2012). These simulations agreed with evolutionary data better than the SS model, although the agreement is slightly worse than in the simulations without indels (data not shown). We note that while indels are important sources of variation for *Drosophila* non-coding sequence (Sinha and Siggia 2005; Nourmohammad and Lässig 2011), the statistical summaries of evolution that we used here focus only on aligned sites, therefore the goodness of fit is not expected to be sensitive to such sources of variation.

4.3 Summary and conclusions

We described here a principled approach to understanding binding site evolution at a higher resolution than previous studies. A seemingly surprising finding of comparative genomics is the unexpected degree of evolutionary flux in regulatory sequences (Dermitzakis and Clark 2002; Balhoff and Wray 2005; Moses et al. 2006). For instance, Emberly et al. (2003) noted that known binding sites in functional CRMs are not much more conserved (between two *Drosophila* species) than in sequences randomly sampled from the genome. A similar exercise of recording the extent to which TF-binding sites are conserved at varying evolutionary distances was conducted more comprehensively by Kim et al. (2009). The conclusion from that study was that sites are lost at a roughly constant rate, that is, the number of site losses is proportional to evolutionary divergence, as might be expected in the absence of lineage-specific selection. ("Site loss" was defined relative to one reference species rather than the common ancestor and, for technical reasons, the study examined losses only.) However, cataloging of site-level evolutionary changes does not address the more fundamental questions: Is the observed rate of site loss lower or greater than expected? What *is* the expected rate? Is there a better way to define this expectation than to base it on random genomic segments (one extreme) or to presume that a functional binding site must remain a binding site, that is, match the TF motif (other extreme)? Why is the site-loss rate for one TF different from another TF? These are the questions that we hope to begin answering with our work. We link the expected evolutionary flux on a TF's binding site to our understanding of that site's function. For this purpose we take

recourse to the regulatory system where current understanding of "cis-regulatory logic", that is, the roles of various binding sites, is among the most advanced – the AP patterning system in the fruitfly embryo (Segal et al. 2008; He et al. 2010). A state-of-the-art computational model of a CRM's regulatory function is coupled with evolutionary simulations under mutation and selection, and the evolutionary histories of a binding site under repeated simulations are used to define the expected rates of site loss and conservation. These expectations agree by and large with the observed rates. Moreover, to a first approximation this approach also explains why sites of one TF evolve at a different rate from those of another TF, although there is room for improvement in this regard. An important aspect of our approach is to assert, in the evolutionary simulations, that the fitness effect of an in-site mutation is context-dependent; put simply, what a mutation does to a site depends on what other sites are nearby. We demonstrate that explicitly modeling this reasonable assertion leads to a better quantitative explanation of binding site evolution.

Simulation frameworks for CRM evolution have recently been proposed in at least two different studies (Lusk and Eisen 2010; He et al. 2012). In both of these studies, the goal was to explain *features of CRM architecture* (e.g., proximity constraints on pairs of sites (Lusk and Eisen 2010) or homotypic clustering of sites (He et al. 2012)) by using a model of CRM function with evolutionary simulations. Our methodology is similar in spirit to Lusk and Eisen (2010), though our goal is to explain *features of CRM evolution under purifying selection*, a fundamentally different goal.

Other frameworks to model the evolution of the regulatory machinery using simulations include (Francois et al. 2007; Cooper et al. 2009; Pujato et al. 2013), all of which study evolution at the gene-regulatory network level. Also noteworthy is the study in Stewart et al. (2012), where a population genetics framework is used to explain the emergence of cooperative binding in regulatory systems, and (Josephides and Moses 2011), where a maximum parsimony approach is used to enumerate all maximally parsimonious evolutionary paths from an inferred ancestral to the current known sequence in *S. cerevisiae*. However, our model is fundamentally different from those models in its resolution: we model evolution at the sequence level, while

the aforementioned studies model evolution at a higher level, with the exception of Stewart et al. (2012), where a sequence simulation was used mainly to validate the population genetics model. At the same time, our model may be used in conjunction with some of the above approaches in future studies.

We attempted to model patterns of binding site conservation and turnover under purifying selection on the CRM's expression readout. This is in contrast to studies that considered a collection of binding sites as evolving under an energy-dependent fitness model (Mustonen and Lässig 2005; Doniger and Fay 2007; Kim et al. 2009), and were concerned primarily with quantifying the average strength of purifying selection on the collection of sites. A similar approach to testing for purifying selection was utilized by Moses (2009). He et al. (2011) recently noted that these approaches are not ideal for detecting positive selection on binding sites, and they examined patterns of polymorphism and divergence in two closely related species (*D. melanogaster* and *D. simulans*) to test for signatures of selection. They found functional site evolution to be primarily under purifying selection. Our study is consistent with this – we found patterns of site conservation (Figure 4.1 and Figure 4.2) in closely related species to be well explained by our simulations, which only implement purifying selection. They also presented evidence for positive selection for both gains and losses of binding sites.

The results presented so far showcase the power of our model for the relatively short evolutionary divergence between *D. melanogaster* and *D. yakuba* (about 13-17 million years (Satta et al. 1987; Satta and Takahata 1990)). However, our methods are also relatively accurate at longer timescales, as will be discussed in Chapter 6.

5 Detecting evolutionary artifacts

5.1 Introduction

Enhancers involved in metazoan development have been known to harbor multiple binding sites for the same transcription factor, a phenomenon known as “homotypic clustering”. This has been documented in invertebrate (Berman et al. 2002; Markstein et al. 2002; Li et al. 2007b) and vertebrate (Sinha et al. 2008; Gotea et al. 2010) genomes alike, and is the basis for several genome-wide enhancer prediction tools (Berman et al. 2002; Markstein et al. 2002; Lifanov et al. 2003; Sinha et al. 2008; Gotea et al. 2010). Several explanations have been offered for this common empirical observation. The common explanation is that multiple homotypic sites in an enhancer (or promoter) are required for the enhancer’s transcriptional efficacy, the desired gene expression levels and ultimately for organismal fitness (Sauer et al. 1995; Hertel et al. 1997). That is, the observed site multiplicity is ostensibly due to selective forces (e.g., (Shultzaberger et al. 2010)). For example, various theories have proposed that site clusters may (a) facilitate lateral diffusion of transcription factor molecules along the DNA, thereby increasing the effective protein concentration (Kim et al. 1987; Coleman and Pugh 1995), or (b) increase occupancy non-linearly through cooperative interactions among sites (Giniger and Ptashne 1988; Hertel et al. 1997) or through simultaneous interaction with the basal transcriptional machinery (Lin et al. 1990; Anderson and Freytag 1991; He et al. 2010). Indeed, non-linear transcriptional response to protein concentration is believed to be important for various phenotypes (Porcher and Dostatni 2010), again suggesting that homotypic clustering may be common due to a selective advantage.

However, common features observed in a class of genomic elements may not be due to functional constraints alone; they may also result from properties of the *fitness landscape* (Mustonen et al. 2008), and from evolutionary sampling of this landscape (Lusk and Eisen 2010). (The space of all possible nucleotide sequences, i.e., genotypes, with a fitness value assigned to every genotype, is henceforth called the fitness landscape.) We hypothesized that the fitness landscape and its evolutionary sampling play an important role in the origin of

homotypic clustering. For instance, a “simple” sequence with one or two perfect binding sites and a “complex” sequence with a number of weaker sites may be equally effective at activating a gene, but complex sequences may be far more abundant and thus favored by evolution. Here, we explore the evolutionary origins of homotypic site clusters in enhancers, through direct examination of the fitness landscape and by simulating the evolution of a simple enhancer. We find that evolution favors complex genotypes even when simpler (more parsimonious) genotypes of comparable fitness exist. This is largely because the space of fit genotypes has more of the former than the latter. Our findings are consistent with an empirical analysis of binding site multiplicities in experimentally characterized enhancers in *D. melanogaster*.

Our results caution against “evolutionary mirages” (Lusk and Eisen 2010), where properties of the evolutionary process lead to genotypic properties that may appear to have mechanistic origins. In particular, they suggest an evolutionary “null hypothesis” for the phenomenon of homotypic clustering, against which alternative explanations, mechanistic or evolutionary, may be assessed in the future.

5.2 Results

5.2.1 Fitness landscape and evolutionary simulations

We begin with the gene expression pattern that constitutes the phenotype for this study. Our enhancers will harbor binding sites for a single transcription factor (TF), whose concentration has an exponentially decaying pattern along an axis (Figure 5.1 A). This mimics the concentration gradient of the morphogen *Bicoid* along the anterior-posterior (A/P) axis of the blastoderm-stage *Drosophila* embryo, but more generally, it reflects the fact that most TFs have spatial/temporal variability in their concentration. The binding specificity of the TF is assumed to be described by the *Bicoid* position weight matrix (Figure 5.1 A and (Bergman et al. 2005)). The expression pattern that a functional enhancer is required to encode is chosen to be identical in shape to the TF's pattern, with 100-fold activation at the highest levels of the TF. (Note that to implement such a linear “readout” of the TF's concentration gradient, an enhancer does not require cooperative interactions among its binding sites.) We next define a

"fitness functional" (denoted by *F*) for any enhancer as a measure of how similar its induced expression profile is to the required profile. This is a number between 0 and 1 (with 1 indicating identity), and is derived from the "Pattern Generating Potential" score in Kazemian et al. (2010) (Figure 5.1 B and Methods.).

An important component of our framework is the quantitative model that maps the enhancer sequence, along with the transcription factor's concentration and binding specificity, to gene expression. We use a model based on statistical thermodynamics that is very similar to that proposed by Shea and Ackers (1985), and discussed and refined by several recent studies (Buchler et al. 2003; Gertz et al. 2009; He et al. 2010). This model has been demonstrated to explain well the spatial patterning of early developmental genes in Drosophila (Zinzen et al. 2006; Segal et al. 2008; He et al. 2010). A brief description of the model is provided in Methods (also Figure 5.1 C), while details can be found in our earlier work (He et al. 2010). Importantly, even weak binding sites contribute to regulation in this model, thus allowing a prediction of the readout encoded by any sequence in the genotype space (and not just those with one or more sites above some threshold). Also, cooperative DNA binding of multiple TF molecules was excluded from the model, for reasons given later (Discussion). We examined the space of 500 bp long sequences (genotypes), their respective expression patterns (phenotypes) and the fitness functional values computed from the phenotypes (Figure 5.1 D).

To characterize the distribution of fit genotypes that an evolutionary process would encounter, we performed Wright-Fisher simulations of a fixed-size population, where each individual is an enhancer genotype. Repeated rounds of random mutation and natural selection were applied to the evolving population, where strength of selection depends on the phenotype (expression pattern) of a sequence and its fitness. (See Methods for details and justification of evolutionary and biophysical parameters.)

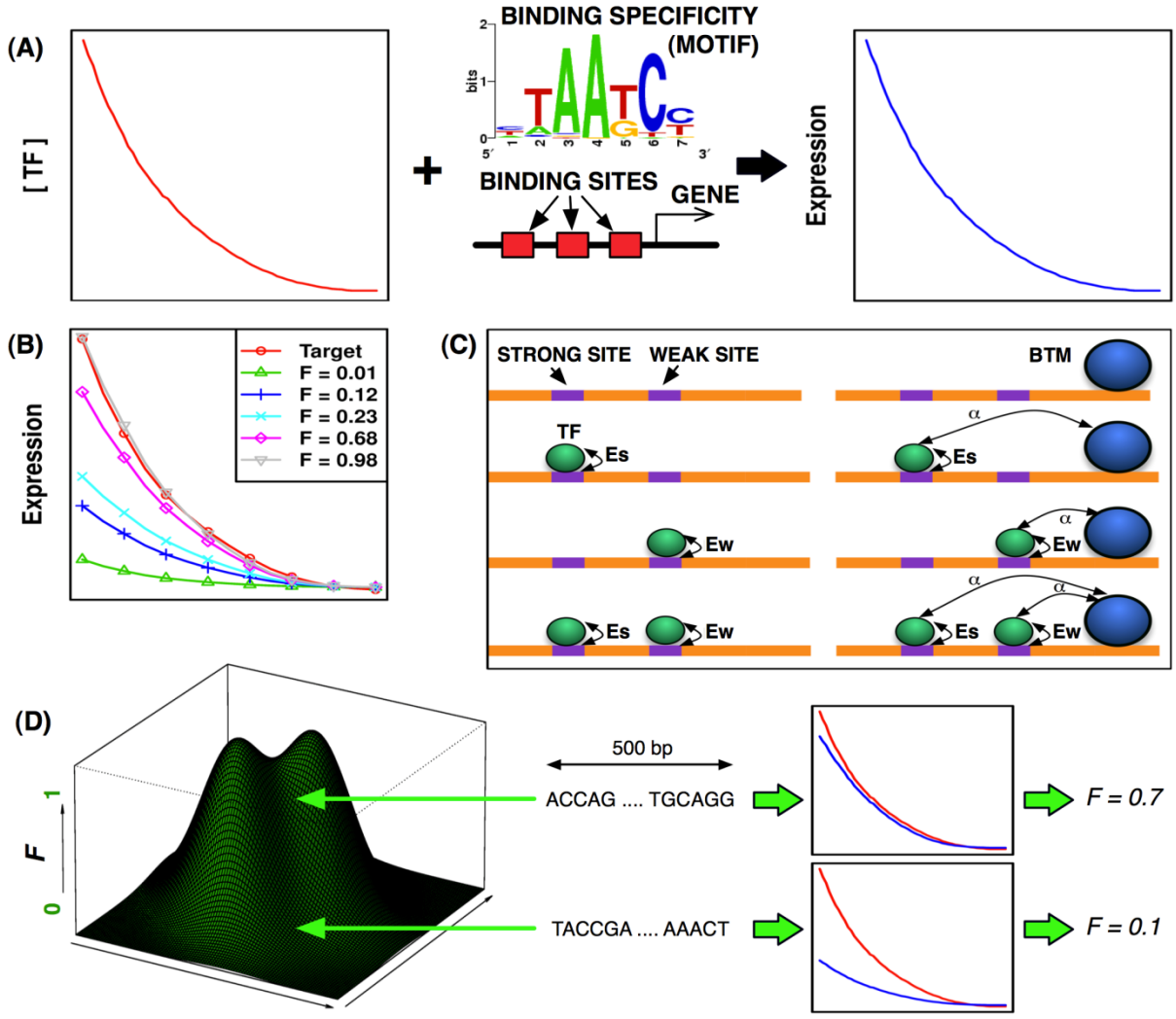


Figure 5.1: A model system for studying evolution of enhancers. (A) The spatial expression pattern of transcription factor (left panel, TF concentration plotted along the anterior-posterior axis) is read by an enhancer (middle, bottom) with sites matching the TF's motif (middle, top), and the result is a spatial expression pattern of the gene regulated by the enhancer (right panel). **(B)** Example gene expression profiles compared to the target profile (red), and associated fitness functional (*F*) values. **(C)** A thermodynamic model of enhancer function. Shown is a sequence with two binding sites (one strong, one weak), which may exist in eight possible configurations of TF molecules (green circles) bound to these sites, and in four of which the basal transcription machinery (BTM) is bound to the promoter. The terms E_s and E_w represent the energetic interactions between a TF molecule and its site, and arrows labeled α denote interactions between TF molecules and BTM. Transcription is assumed to be initiated only when BTM is bound; thus the total probability of the four configurations on the right determines the activation level of the gene due to this enhancer (see Methods for details) **(D)** A cartoon illustration of the fitness landscape. All possible sequences are points on the horizontal plane. Each sequence corresponds to an expression pattern, which determines its fitness functional (*F*) value.

5.2.2 Sampling by evolution shows abundance of complex genotypes

Fifty independent evolutionary simulations were run for 10^6 generations each; adaptation typically happened within 10^5 generations, and the average fitness functional for the population stayed above $F = 0.8$ thereafter (Figure 5.2 A). We sampled *post*-adaptation genotypes from all simulations and examined the site multiplicity of this evolutionary sample of fit genotypes (Figure 5.2 B). (Site multiplicity is the number of binding sites in the genotype, defined by a threshold on their binding affinity relative to that of the optimal site. The threshold used here is 0.25 times the binding affinity of a perfect site; see Methods for details). While the most parsimonious genotypes sampled use only 1 above-threshold site, the mode of the distribution is at 5 sites, clearly demonstrating that evolutionary sampling favors genotypes with relatively high site multiplicity. This observed bias is not due to the complex genotypes in the pool having higher fitness (Figure 5.2 C). At the same time, *very* complex genotypes (e.g., those with > 7 sites at the threshold) are also rare in the evolutionary sample. The trends of Figure 5.2 B are also seen when defining sites with a stricter threshold of relative affinity ≥ 0.5 (Figure 5.2 D). We also plotted the genotype frequency at different values of the "occupancy" of the TF on the entire enhancer (Figure 5.2 E). (Occupancy is defined by the thermodynamic model as the average number of sites bound by TF molecules, and is independent of any threshold on site affinity; see Methods.) Clearly, the range of observed occupancy values is much smaller than the ranges of site multiplicity (Figure 5.2 B,D). In other words, selection ensures that the genotypes sampled after adaptation lie in a narrow range of occupancy (which is closely related to fitness); however, the same occupancy level (and hence fitness) can be achieved through a wide range of site multiplicities.

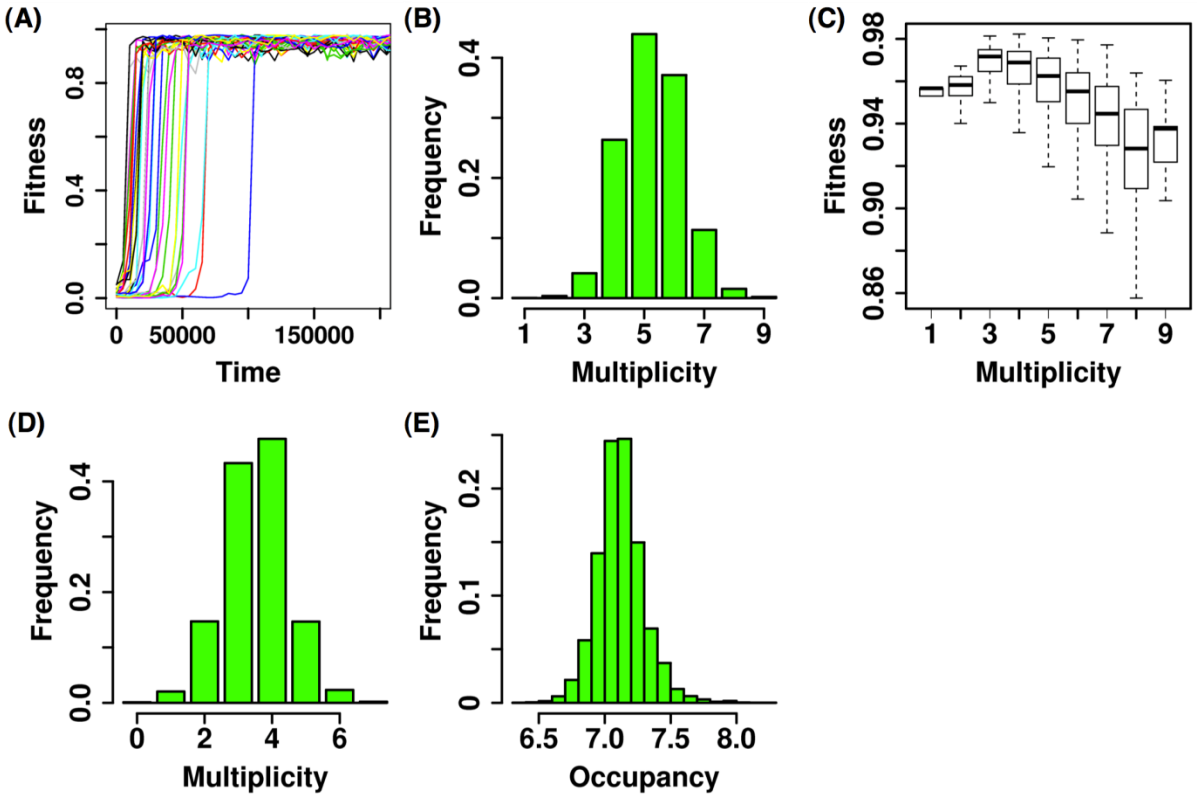


Figure 5.2: Results of evolutionary simulations. (A) Time series of (average) population fitness, showing adaptation. Each curve represents the history of one population (truncated at 200,000 generations). (B) The distribution of site multiplicity (number of binding sites) in post-adaptation genotypes (Five random individuals with $F \geq 0.8$ were sampled from the population every 5000 generations). The X-axis is the number of sites (at relative affinity ≥ 0.25) and the Y-axis is the frequency of genotypes with that multiplicity. (C) Box plot of fitness (F) values of genotypes with different site multiplicities. (D) Same as plot (B), but for a higher affinity threshold, 0.50. (E) Distribution of TF occupancy in post-adaptation genotypes.

5.2.3 Causes of the evolutionary bias towards complex genotypes

Abundance of complex genotypes in the fitness landscape: The distribution of genotypes sampled by evolution is shaped, to a large extent, by the fitness landscape. (For example, see (Sella and Hirsh 2005) for an expression for the equilibrium probability of sampling a genotype, as a function of its fitness functional *F.*) Thus, a possible explanation for the complex genotype bias seen above is that the fitness landscape has a relative abundance of such genotypes, at high *F* values.

We therefore examined the fitness landscape directly, with the goal of characterizing the site multiplicity of all fit genotypes. First, we analytically estimated the frequency of genotypes with exactly *k* binding sites at relative affinity ≥ 0.25 (Methods), shown in Figure 5.3 A. We see that for the most part, genotypes with fewer sites are more abundant. Next, for each *k*, we sampled genotypes with exactly *k* binding sites (uniformly at random), computed the fitness functional for each genotype, and thus estimated the probability that such genotypes are fit ($F \geq 0.8$) (Figure 5.3 B). Finally, multiplying the quantities shown in Figure 5.3 (A) and 3(B), we obtained the relative proportion of *k*-site genotypes in the space of fit genotypes (Figure 5.3 C): the three most abundant site-multiplicity values are *k*=6,5,7, in that order, together accounting for about 90% of the total frequency. Thus, complex genotypes are indeed more common among all fit genotypes and this explains their dominance in the results of evolutionary simulation. We also noted clear examples of how one genotype class can be evolutionarily preferred over another class (e.g., *5*-site vs. *7*-site genotypes, Figure 5.2 B) due to greater frequency (Figure 5.3 C), despite being less fit on average (Figure 5.3 B).

Importance of weak binding sites: We next analyzed why complex sequences are frequent among fit genotypes. We hypothesized that the strength of binding sites play a major role here: that complex genotypes make use of contributions from many sub-optimal sites to achieve the same net occupancy of the TF on the enhancer as might be achieved through fewer, closer-to-optimal sites. If this is the case, the complex genotype bias in the fitness landscape (Figure 5.3 C) should become less prominent as we make the threshold for counting sites more stringent. We found this to be the case indeed, as shown in Figure 5.3 D. For instance, at the high threshold of relative affinity = 0.8, where only the optimal site gets counted, the mode of the observed distribution (site multiplicity *k*=2) also corresponds to the most parsimonious (simplest) genotype(s) observed to achieve the fitness criterion of $F \geq 0.8$; in other words, the complex genotype bias is not seen. A direct examination of site strengths revealed that complex genotypes have weaker sites on average than simpler genotypes (Figure 5.3 E).

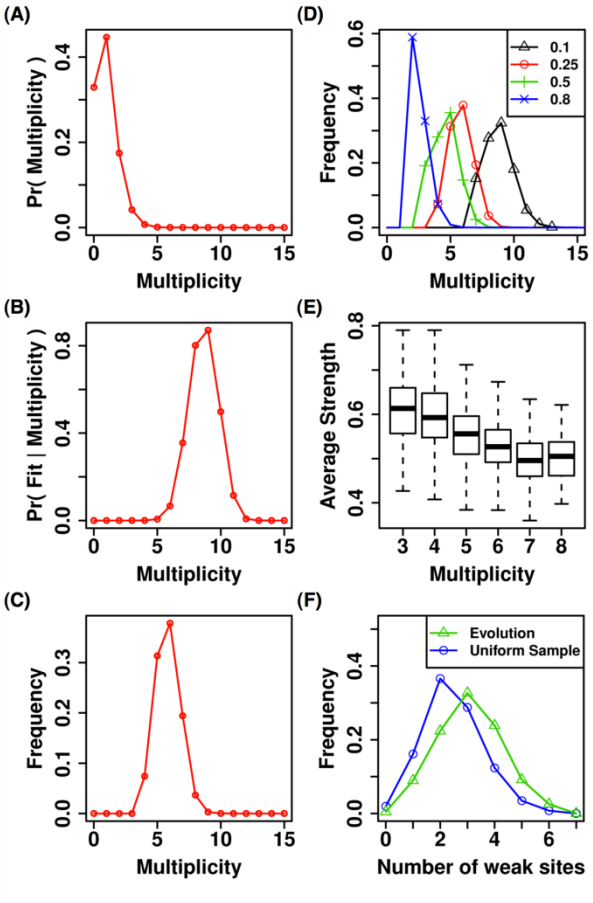


Figure 5.3: Genotype frequency and properties of fit genotypes. (A) Relative frequency of genotypes with different site multiplicities: the number of sequences with *k* binding sites (at relative affinity ≥ 0.25) is estimated analytically, for each value of *k*. (B) Probability of a genotype with *k* sites being fit (F ≥ 0.8). (C) Frequency of *k*-site genotypes among all fit sequences, calculated by multiplying the relative frequency in (A) and the probability of being fit (B). (D) Same as plot (C), at different relative affinity thresholds (0.1, 0.25, 0.5, 0.8). (E) Average relative affinity of binding sites in *k*-site genotypes. (F) Histograms of sub-sites (relative affinity < 0.25), for evolutionary (green) and uniform (blue) samples of genotypes with *k*=6 sites at relative affinity 0.25.

Thus, broadly speaking, there are two types of fit genotypes: simple sequences, with few strong sites, and complex sequences, with more weak sites. Both types of genotypes can achieve high fitness, but complex sequences with weak sites are common in the evolutionary samples (Figure 5.2 B,D) due to their high genotype frequency (Figure 5.3 C). To illustrate the intuition behind this, we present a simple theoretical calculation. Let us characterize a genotype by the two integers (k,m) where *k* is the number of sites, and *m* is the strength of each site (defined here, for simplicity, by the number of mismatches relative to the optimal site). The abundance of *(k,m)* genotypes can be calculated as:

$$N(k,m) = \binom{L}{k} 2^k \left[\binom{l}{m} 3^m \right]^k 4^{L-kl} \qquad (1)$$

where *L* is the length of the enhancer and *l* is the length of each site. Using this formula and the parameter values in our setting (*L* = 500 and *l* = 7), we find that complex sequences can be more common than simple ones (Supplementary Fig. S1 of (He et al. 2012)). For instance, we see that *N(2,1)* is about 13 times larger than *N(1,0)*, i.e., genotypes with two suboptimal sites are 13 times more frequent than genotypes with one optimal site. Similarly, *N(3,1)* is ~6 times larger than *N(1,0)*. If we assume, for instance, that that one optimal site can be functionally replaced by a few suboptimal sites (e.g., 2-3 sites with 1 mismatch each), the class of fit genotypes will have a relative abundance of complex genotypes. This simplistic calculation, which is not tied to the precise genotype-phenotype mapping and its parameters, reveals the main idea behind an evolutionary origin of homotypic site clustering. In the Supplementary text, we explore a different theoretical model of binding sites where certain positions of a binding site, strong or weak, must remain invariant, and we find the same intuitive explanation of HTC to be revealed by this alternative model.

An evolutionary signature: We designate the samples we obtained for studying the properties of the fit genotypes (Figure 5.3 C) as "uniform samples" to distinguish them from evolutionary samples, because the way evolution explores the fitness landscape depends on history (thus not uniform sampling). We noted that the evolutionary samples (Figure 5.2 B) and uniform samples (Figure 5.3 C) of the same population (i.e., all fit genotypes) have similar site multiplicity distributions, with most of their probability mass concentrated on the same values (*k* = 5,6). However, the two distributions also have significant differences, e.g., evolutionary samples include a greater representation of *k*=4 genotypes compared to uniform samples (probability 0.21 vs. 0.07). This particular statistical observation led us to an interesting characterization of evolutionarily sampled genotypes. We first noted that the average fitness of *k*=4 genotypes is comparable to that of *k*=5 genotypes in the evolutionary samples (Figure 5.2 C), but substantially lower in uniform samples (Supplementary Fig. S2 of (He et al. 2012)). In other words, evolution finds only the fittest ($F \approx 0.97$) among all fit *k*=4 genotypes. Investigating this further, we noted that the *k=4* genotypes that evolution finds have unusually many "sub-sites" (sites below the strength threshold used), in addition to the 4 sites, that contribute to the

occupancy and hence to fitness of the enhancer. This is clear from Figure 5.3 F, where histograms of "sub-site" multiplicity show that fit genotypes sampled by evolution (green) and are significantly enriched in sub-sites compared to uniform samples (blue). In other words, evolutionary samples have a greater *spread* of site strengths than random expectation (represented by the uniform samples). This is what leads, in this case, to k=4 genotypes found by evolution having unusually high fitness, and consequently, a higher relative frequency. Interestingly, this evolutionary signature was reported previously, as an abundance of sub-sites near functional binding sites, in a systematic analysis of 11 mammalian genomes (Reid 2007).

5.2.4 Temporal profile of site multiplicity in an evolving enhancer

We noted above (Figure 5.2) that the evolutionary process frequently samples complex genotypes *after* adaptation has been reached. However, it is plausible that parsimonious genotypes serve as the *entry points* to the space of fit genotypes that evolution explores. That is, evolution may be "stumbling into" parsimonious genotypes first because they have few sites, and after one fit genotype has been found, subsequent gain and loss of sites leads to more complex genotypes. We therefore asked if the evolving enhancer is parsimonious at the time of reaching adaptation, and acquires additional sites post-adaptation. Surprisingly, we found this not to be the case. Instead, we observed that the most parsimonious genotype reached by evolution (over a long period) is typically reached post-adaptation. Figure 5.4 A shows three typical simulations, in terms of how the site multiplicity changes with time, before as well as after adaptation. Note that in each simulation, the most parsimonious fit genotype (arrows) is encountered well after adaptation was reached (shown as the point of transition from black to colored lines). This is true of most of our simulations: in 70% of our simulations, the most parsimonious fit genotype had at least two fewer sites than the genotype at which adaptation was reached (Figure 5.4 B).

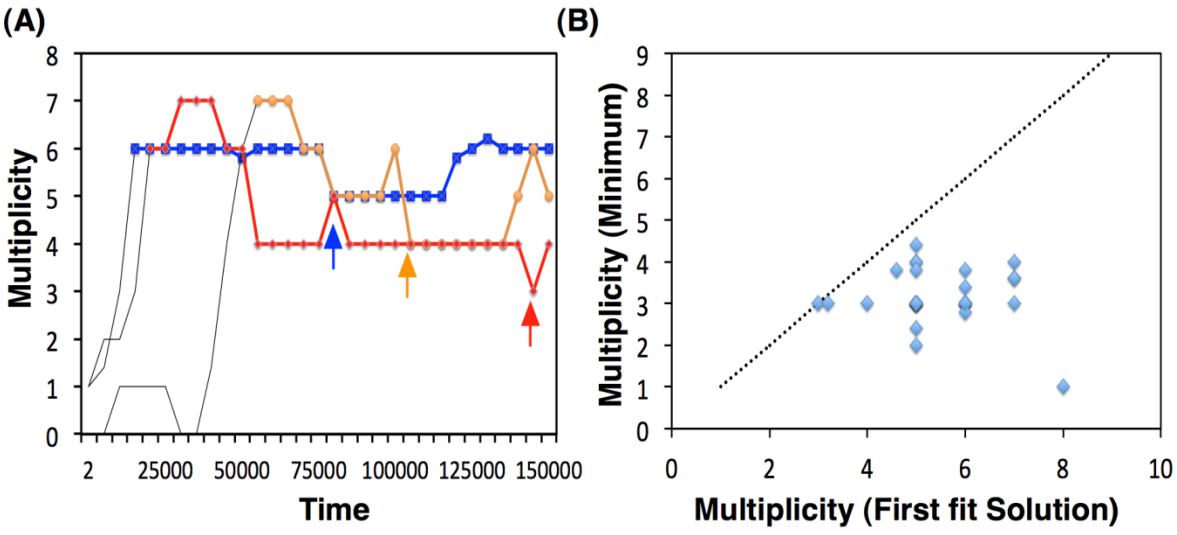


Figure 5.4: Temporal dynamics of site multiplicity. (A) Average site multiplicity of genotypes in the evolving population, as a function of time, for three typical simulations. The color part of each curve indicates post-adaptation profile (F ≥ 0.8); the grey part indicates pre-adaptation. **(B)** Average site multiplicity of genotypes at adaptation (X-axis) vs. the minimum multiplicity encountered post-adaptation (Y-axis). Each point represents one simulation.

5.2.5 Site multiplicity distributions in *Drosophila* enhancers

Our study is based on the motif and concentration profile of the *Bicoid* transcription factor, which activates expression in the anterior half of the blastoderm-stage embryo in *Drosophila*. Therefore, it is instructive to examine if our observations about site multiplicity distributions (in synthetic genotypes) are mirrored in real enhancers as well. We collected 21 *bona fide* enhancers that use *Bicoid* binding sites to drive anterior expression in the early embryo in *D. melanogaster*. For each enhancer, we also collected orthologs from (up to) five other moderately diverged species from the *Drosophila* group, and computed their site multiplicity at the same threshold as in Figure 5.2 B above. These are shown in Figure 5.5 A, grouped by orthology. If one assumes that orthologous enhancers have similar fitness, this plot suggests that variability in site multiplicity and abundance of non-parsimonious genotypes is true of real fitness landscapes. However, orthologous enhancers may vary in their transcriptional outputs, and even if they have the same output, different orthologs may utilize *Bicoid* binding to different extents (for instance, by making use of other transcription factors). Therefore, we next estimated the occupancy of every enhancer in our collection (using the same procedure as in

Figure 5.2 E), and examined the site multiplicity distributions of enhancers grouped by occupancy. (Recall that in our simulations, post-adaptation genotypes exhibit a narrow range of occupancy centered at ~7 (Figure 5.2 E).) The results (Figure 5.5 B) suggest that if we use estimated occupancy as a surrogate for the transcriptional effect of *Bicoid*, evolutionary samples of the fitness landscape exhibit (qualitatively) the same kind of variability of site counts that our theoretical study anticipates. For example, enhancers with estimated occupancy ~7 have site multiplicity in the range 2-5, with the median at 4. (Compare this to artificial evolutionary samples in Figure 5.2 B, mostly with 3-7 sites and a median of 5.) That this is a qualitative rather than quantitative agreement is expected, since the quantitative model used in our simulations almost certainly misses certain aspects of the real enhancers' regulation.

5.2.6 Other potential causes of complex genotype bias

Finally, we investigated additional factors that may influence the emergence of a complex genotype bias in enhancers, including non-equilibrium sampling of the fitness landscape, short local duplications in DNA, and differences in stochasticity of gene expression induced by different genotypes.

Local topography of fitness landscape: The distribution of evolutionarily sampled genotypes (Figure 5.2) may depend on local properties of the fitness landscape. For instance, high fitness genotypes in a relatively "rugged" region may be sampled more or less frequently than similar-fitness genotypes in a smoother region (Weinberger 1991; Smith et al. 2002). We quantified the local ruggedness of the fitness landscape around a genotype by its "average correlation length" (ACL (Hordijk 1995), Supplementary Fig. S9 of (He et al. 2012)), and found that genotypes with k=4 – 8 sites had similar ACL, suggesting that topographical differences, at least to the extent characterized by the ACL score, do not significantly influence the complex genotype bias.

Effect of local duplications: A remarkably high coverage of short tandem repeats has been observed in *Drosophila* enhancers (Sinha and Siggia 2005), suggesting that short local duplications may play an important role in regulatory sequence evolution, and perhaps lead to homotypic site clustering. To investigate this, we compared the results of evolutionary

simulations with substitutions, short insertions and deletions, to simulations where all or part of the insertions were local duplications. However, we observed no difference in either the complex genotype bias or the adaptation time in simulations with or without local duplications (data not shown). Future work will have to examine the role of local duplications in enhancer evolution under varying assumptions about the underlying indel and duplication rates and length distributions.

Noise characteristics of complex genotypes: The binding site composition of an enhancer has the potential to affect intrinsic noise (stochasticity) in gene expression levels and thus the robustness of biological processes (Raser and O'Shea 2004; Kaern et al. 2005). In particular, a recent study (Holloway et al. 2011) shows that greater number or strengths of *Bicoid* sites in the *hunchback* gene promoter leads to reduced noise in *hunchback* expression (while increasing the expression levels). We therefore asked if the (high fitness) genotypes sampled by our evolutionary simulations might reveal a correlation between site multiplicity and noise in gene expression. We estimated variance in TF occupancy on each enhancer (occupancy and expression level are correlated in our model), and found a strong negative correlation with site multiplicity (Supplementary Fig. S10A of (He et al. 2012)). Importantly, this correlation exists despite the mean occupancy being roughly constant (Supplementary Fig. S10B of (He et al. 2012)). Phenotypic consequences of reduced noise in expression may therefore be an important factor leading to the complex genotype bias observed in real enhancers. However, since such consequences were not factored into our fitness function, we conclude that the bias towards homotypic clustering can arise even in the absence of a noise-fitness relationship.

5.3 Methods

5.3.1 Strength of binding sites and TF occupancy

The strength of a binding site is defined as its binding affinity relative to the strongest ("consensus") site. It is a number between 0 and 1, and a site with a relative affinity of 0.1, for example, is 10 times weaker than the optimal site, in terms of association constant. Let LLR(s) be the log likelihood ratio score of site s, computed based on the known position weight matrix

(PWM) of the TF and the background nucleotide distribution (Stormo 2000). The site's strength (relative affinity) is computed as *exp(LLR(s)—LLR(s_{opt})),* where s_{opt} is the optimal site.

The TF's occupancy at an enhancer is defined as the sum of the fractional occupancy of all sites in the enhancer, at maximum TF concentration, as computed by the GEMSTAT model. Fractional occupancy of a site is given by the total statistical weight of all configurations where the site is bound, relative to that of all configurations.

5.3.2 Analytical estimation of number of genotypes with k sites

The relative affinity threshold is converted to a p-value *p* of the site LLR, and the (relative) number of genotypes with at least *k* sites at this threshold is computed as $\binom{L}{k} 2^k p^k$, where *L* is the length of the enhancer sequence. Taking differences between successive values of *k* gives the desired number of genotypes, up to a constant of proportionality.

5.3.3 Parameterization of the expression model and evolutionary simulation

The two main parameters of the GEMSTAT model are the "DNA binding" parameter (β) and the "activation strength" parameter (α). Our default parameter settings were $\beta = 5$, $\alpha = 2$. These values were obtained from a separate exercise where we simultaneously modeled the expression profiles of 20 AP axis patterning enhancers (and the non-expression of equally many random sequences), using the binding specificities (motifs) of six different TFs (see Supplementary Fig. S4 of (He et al. 2012)). The *Bicoid* transcription factor, whose motif and concentration profile we have used throughout our study, was assigned the above values (approximately) in the trained model. To get some intuition into what these values mean, we note that $\beta = 5$ implies that the consensus binding site for the TF has a fractional occupancy of 5/6 at maximum TF concentration. Likewise, $\alpha = 2$ implies that a site with fractional occupancy ≈ 1 induces 2-fold activation of gene expression, and under our settings for synergistic activation, about ~7 high occupancy sites are needed to achieve 100-fold activation. This number is roughly consistent with the number of Bicoid sites found in the well-studied hunchback promoter, that drives anterior expression. Furthermore, a simple calculation shows

that with these parameter settings, a random sequence of length 500 bp is expected to show no expression. Thus, we believe that our default settings for the thermodynamic model parameters are realistic. We also repeated our simulations with an alternative setting ($\beta = 1$, $\alpha = 5$, Supplementary Fig. S5 of (He et al. 2012)), and found little difference in the main observations reported above.

The key parameters in the evolutionary simulations are the population size ($2N$), the mutation rate per nucleotide per generation (μ) (or equivalently, $2N\mu$) and the selection coefficient (s) (or equivalently, $4Ns$). Default settings of these parameters were $2N = 100$, $2N\mu = 10^{-3}$, and $4Ns = 100$. Standard values of the population size and mutation rate, from the literature, are $2N \sim 10^5 - 10^6$ (Thornton and Andolfatto 2006) and $\mu \sim 10^{-8} - 10^{-9}$ (Drake et al. 1998) giving us $2N\mu$ in the range of 0.01 – 0.0001, which is approximately what we set it to be. We used time rescaling (Hoggart et al. 2007) to speed up our simulations. Here, the population size is scaled down by a constant (we used $\lambda = 1000$), keeping $2N\mu$ and $4Ns$ unchanged; t generations of simulation in this scheme is approximately equivalent to λt generations of simulation in the absence of rescaling. Thus, the default setting of $2N=100$ is equivalent to $2N=10^5$ without time scaling. We repeated the simulations with a larger population size of $2N=1000$ (equivalent to $2N=10^6$, unscaled) and noted that the observed trends were unchanged (Supplementary Fig. S6 of (He et al. 2012)). We set the selection coefficient s of a genotype as $s = FK$, where $F \in [0,1]$ is the fitness functional of the genotype, and K is the selection coefficient of the fittest genotype ($F=1$) in relation to the least fit genotype ($F=0$). The latter was set to a value of $50 \cdot 1/2N$ by default, indicating strong selection ($2NK = 50$). Note that in any one generation, there is a relatively small difference in F between the fittest genotype and the wild type; this means that the effective selection coefficient s for the fittest genotype is typically much smaller than $50/2N$. We also repeated our simulations with $2NK$ set to 10 and 20. Adaptation was often not observed in the sampled time at the former value, hence the corresponding results are not shown. Results of simulations with $2NK = 20$ are shown in Supplementary Fig. S7 of (He et al. 2012) and support our claims above. All simulations were performed in the absence of insertions and deletions, which have been suggested as important influences in the

evolutionary dynamics of regulatory sequences (Sinha and Siggia 2005; Lusk and Eisen 2010). While a detailed examination of this influence was not pursued here, we repeated our simulations with indels (at rates proposed in the literature) and found our observations about distributions of site multiplicity to be unchanged (Supplementary Fig. S8 of (He et al. 2012)).

5.3.4 Drosophila enhancers

We collected 21 Bicoid-driven enhancers from *D. melanogaster* with functions in anterior-posterior patterning (Ochoa-Espinosa et al. 2005; Halfon et al. 2008b). Orthologous sequences from five other species in the melanogaster group (*D. ananassae, D. pseudoobscura, D. virilis, D. mojavensis*, and *D. grimshawi*) were extracted using the liftover tool http://genome.ucsc.edu/cgi-bin/hgLiftOver.

5.4 Summary and conclusions

A fundamental aspect of understanding the complexity and design of a biological system is whether these features are functional requirements, or consequences of the evolutionary process (Lynch 2007a). For instance, a complex design may be chosen by evolution not because of any inherent functional advantages over alternative designs, but because it is more easily found by evolution (Soyer and Bonhoeffer 2006). We studied this question in the context of *cis*-regulatory sequences. The design feature we investigated is the homotypic clustering (HTC) of transcription factor binding sites, found in regulatory sequences across major animal kingdoms (Lifanov et al. 2003; Gotea et al. 2010). The relative simplicity of the system we studied, where the phenotype (expression pattern) of a sequence can be defined using a well-studied biophysical model, allows us to simulate its evolution and perform controlled analysis. Our results show that even when simpler designs exist for the desired expression pattern, relatively complex designs (genotypes with more sites) are more readily reached by evolution (Figure 5.2 B,D). This is, to a large extent, because those complex sequences occupy a larger proportion of the space of fit genotypes (Figure 5.3 C). There are more ways to “build” a fit enhancer with many weak sites than with a few strong sites, and this is why evolution finds the former type more often. We also observed a subtle but clear evolutionary signature in the synthetic enhancers: evolutionary samples tend to have a broader spread of site strengths (Figure 5.3 F)

than expected from a uniform sampling of all fit genotypes. We explored the temporal profiles of site multiplicity in an evolving enhancer, and found, somewhat surprisingly, that simpler designs are not necessarily the precursors of more complex designs that evolve post adaptation (Figure 5.4). We examined site multiplicities of *Bicoid*-driven enhancers in Drosophila species, and found a characteristically broad range of multiplicities among enhancers grouped by orthology or by estimated *Bicoid* occupancy (Figure 5.5), providing empirical evidence for the complex genotype bias we observe in simulations. Finally, we investigated alternative sources of this bias, and found that local topography of the fitness landscape (around a fit genotype) does not play a significant role, nor does the phenomenon of short local duplications in the sequence, at least within the parameter ranges we explored. On the other hand, the higher fidelity (reduced noise in gene expression) associated with complex genotypes *is* a potential cause of their relative abundance, even though we did not explicitly demonstrate this within our simulation framework.

We note that to an extent, HTC does arise from functional requirements – if an enhancer driving the appropriate expression level requires an occupancy of say 5, it must harbor at least five sites; this is a functional constraint. At the same time, it is accepted that multiple weak sites may function as well as one or few strong binding sites (Roider et al. 2007; Shultzaberger et al. 2010), suggesting that the neutral space (Wagner 2007) of fit genotypes may be highly diverse. We propose that this diversity is a key determinant of enhancer composition, and that the required TF occupancy is more likely to be implemented through a greater number of sites (including sub-optimal ones) than with the minimal number of optimal sites.

Earlier work has proposed specific mechanistic explanations of HTC: that multiple sites may facilitate TF-DNA interaction synergistically (Giniger and Ptashne 1988; Lin et al. 1990; Anderson and Freytag 1991; Hertel et al. 1997; He et al. 2010), or that HTC can make sequences more robust to genetic and environmental perturbations (Ludwig et al. 1998), among others (Gotea et al. 2010). However, our simulations clearly showed a complex genotype bias even in the absence of cooperative interactions between sites (also see Supplementary Fig. S3 of (He et al. 2012)), and despite the fact that our fitness function does not incorporate robustness. Thus, we

offer a plausible explanation for HTC that relies upon fairly general assumptions about the underlying biochemical model and fitness function. This provides a baseline that more specific mechanistic explanations may be compared to, or used in conjunction with. We do note that our results rely upon contributions of multiple sites being free from spatial constraints, unlike what is proposed in enhanceosomal models of enhancer function (Arnosti and Kulkarni 2005). Without this assumption, calculation of the abundance of genotypes may favor simple instead of complex sequences. Many studies to date have found the arrangement of binding sites in metazoan enhancers to be extremely flexible (Brown et al. 2007; He et al. 2009; Liberman and Stathopoulos 2009), supporting our assumption, but this issue is currently open to debate.

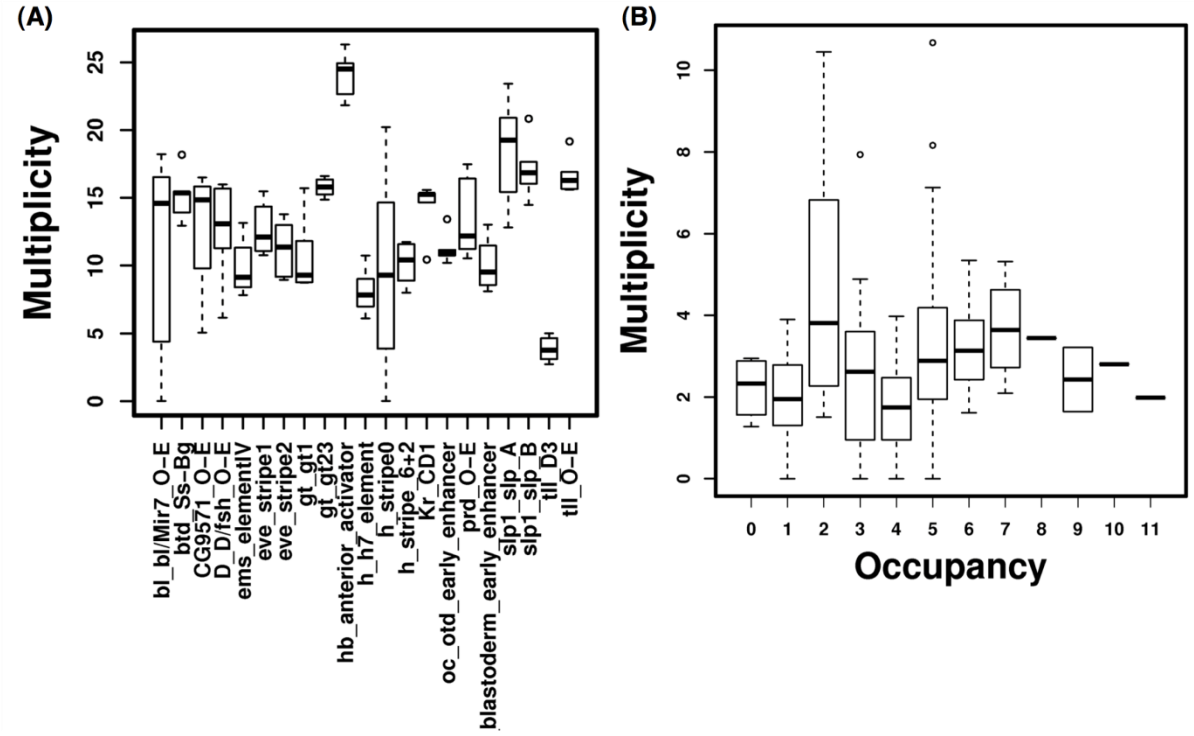


Figure 5.5: Multiplicity and occupancy of orthologous enhancers of Drosophila. (A) Box plot of site multiplicities (at relative affinity ≥ 0.25) for orthologs of *Bicoid*-driven enhancers. Multiplicity values are not normalized for length, since each orthology group has relatively little length variation. **(B)** Box plot of site multiplicities (y-axis) for *Bicoid*-driven enhancers grouped by TF occupancy (x-axis). Occupancy and multiplicity values are normalized by length, since each value of occupancy includes enhancers of widely different lengths.

Our intuitive explanation of HTC is based on the assumption that the function of a strong binding site can be replaced by multiple weak ones. However, there are many reported cases where an enhancer may harbor multiple high-affinity binding sites. We hypothesize several possible explanations: for instance, some enhancers may demand a high level of TF affinity that requires multiple high-affinity sites; the enhanceosome model as explained above makes it impossible to trade one strong site for multiple weak ones. Also, non-adaptive forces such as short tandem duplication may facilitate the occurrence of multiple high-affinity sites.

A recent study (Paixao and Azevedo 2010) examines the multiplicity of binding sites in enhancers, and uses simulations to show that this is largely due to recombination and weak direct selection for multiplicity. However, the definition of multiplicity (as the presence of two or more perfect binding sites) by Paixao et al. is very different from our definition, making its central question distinct from ours. Khatri et al. (Khatri et al. 2009) studied the evolution of enhancer sequences using a model system similar to ours, but focused on the question of whether the optimal phenotype is reached (or not), in an adaptive process.

One way to interpret our results is that in genotypes found by evolution, the desired function (phenotype) is distributed into multiple weak components, instead of being concentrated on one or two strong ones. Such “distributed” designs, if allowed, may be a common feature of other systems. For example, signal transduction processes are often characterized by a long cascade of signaling events, where each step may serve only a small piece of the overall function of the pathway (e.g., extent of signal amplification) (Li and Qian 2003; Soyer and Bonhoeffer 2006). Our analysis suggests that a distributed design may in fact be a consequence of the evolutionary process, where both fitness and abundance of genotypes are important determining factors for the sampled designs.

6 Understanding the mechanisms of gene regulation

6.1 Background and motivation

In Chapter 4 we demonstrated how our approach could be used to model evolutionary data accurately when modeling changes between the relatively close species *D. melanogaster* and *D. yakuba* (divergence of about 13-17 million years (Satta et al. 1987; Satta and Takahata 1990)). We were also interested in modeling evolutionary data representing longer time spans, and also in exploring different summary statistics of binding site-evolution. To this effect, we tried to model the site loss rate reported by Kim et al. (Kim et al. 2009) for each of five different TFs in the AP patterning CRMs noted above. Kim et al. (Kim et al. 2009) noticed that the loss of binding sites between *D. melanogaster* and the 11 other species could be well explained using linear regression against evolutionary divergence from *D. melanogaster* (R^2 of at least 0.91 (for HB) and as high as 0.99 (for BCD)). This linear decrease in the fraction of shared binding sites is an indication of a molecular clock, a phenomenon usually taken as an evidence for the lack of lineage specific selection. Kim et al. (Kim et al. 2009) also reported the loss rate for each specific TF, pointing out that this loss rate is lower from what would be expected by chance, as might be expected from a collection of functional sites. Therefore, the loss rate should represent a good estimation of the selective pressure for conservation of binding sites for specific TFs.

6.2 Results

6.2.1 Simulating evolution across larger evolutionary distances

In the previous sections we attempted to explain evolutionary data summarized in the form of energy difference histograms for orthologous pairs of sites in two closely related species (*D. melanogaster* and *D. yakuba*), where most strong sites in one species are retained as strong sites in the other species. Kim et al. (2009) proposed a complementary method to describe binding site evolution, which is geared towards larger evolutionary spans. Using *D. melanogaster* as a reference species, they counted what percentage of predicted sites of a given TF is “lost” in a second *Drosophila* species.

To estimate the site loss rates for each TF, we first created a multiple alignment of each CRM and adjusted it locally following the procedure described in the previous section and in (Kim et al. 2009). Next, we predicted binding sites in the *D. mel* CRM based on the TF's PWM and an LLR p-value threshold. We calculated the number of sites (predicted by the same method) in each species, counting only sites that are aligned to the collection of sites in *D. mel,* and plotted that number as a function of evolutionary distance. Next, we fit a linear model to this data and use the slope of the resulting line to calculate the loss rate. Here, a site loss was called if the *D. melanogaster* site was partly or entirely deleted in the second species, or had accumulated mutations that reduce its predicted binding affinity below the defining threshold. The site loss percentage thus computed was plotted for different choices of the second species, revealing that this percentage varies linearly with divergence time. Our next tests of evolutionary models deal with this alternative summarization of evolutionary data.

We performed PEBCRES simulations of CRM evolution for a fixed number of generations that matches the evolutionary distance between *D. melanogaster* and *D. willistoni* (see Methods), and recorded the site loss percentage between *D. melanogaster* and each of 11 other *Drosophila* species – *D. simulans*, *D. sechellia*, *D. erecta*, *D. yakuba, D. ananassae, D. pseudoobscura, D. persimilis, D. virilis, D. grimshawi, D. mojavensis and D. willistoni*, with *D. willistoni* representing the greatest divergence and *D. simulans* representing the least divergence. This "site loss profile" was computed for each of five different TFs, examining sites over the same 37 *D. melanogaster* CRMs analyzed in previous sections. Each TF's site loss profile was compared to the analogous profile obtained from alignments of the 37 *D. melanogaster* CRMs with orthologous CRMs in the 11 other species, as in Kim et al. (2009). Figure 6.1 (A) and (B) show the site loss profile for the TF BCD, from PEBCRES simulations and real data respectively. The first thing to note in both profiles is that the percentage of *D. melanogaster* sites lost in a second species increases linearly (R^2 of 1.00 and 0.97 respectively) with the evolutionary divergence between *D. melanogaster* and that species. Observing such a "molecular clock" in evolutionary data is often taken as evidence against species (or branch) specific adaptive evolution, and indicates that the collection of sites analyzed evolved

predominantly under purifying selection. Indeed, this was the interpretation offered by Kim et al. (2009). In our simulations, the observation of a molecular clock is trivial since the model imposes no branch-specific selection. However, the slope of the linear relationship, which we call the "loss rate", may be treated as a summary statistic to be compared between model and data. Thus, in Figure 6.1 (A) and (B), the loss rate of 0.15 from real data is well matched to the value of 0.18 observed in PEBCRES simulations. To our knowledge, this is the first attempt to quantitatively explain the rate of binding site loss or gain with models of sequence function and evolution. Note that we only examine site loss rates here (and not gains), for the same technical reasons encountered by Kim et al. (2009): a recorded site loss is a more reliable observation, while site gains are more likely to be conflated with spurious site predictions.

To better illustrate the agreement between loss rates from model and data, we devised the representation scheme shown in Figure 6.1 (C), where each TF is represented by a rectangle. The x and y axes of the plot represent the loss rate inferred from model simulations and real data respectively. The center of the rectangle (marked by a cross) represents the respective loss rates from the procedure outlined above, that is, from an examination of sites in all 37 CRMs included in our analysis. The sides of the rectangle represent an error estimate as calculated by a resampling procedure using 50 samples of 18 CRMs each (out of the full set of 37) for real data and 50 samples of 500 CRMs each (~10-15 simulations per CRM) for model predictions. The diagonal line represents perfect agreement between data and model. All five TFs whose sites were examined are represented on this plot. We find the model-based loss rates to agree with real loss rates for four out of five TFs, with the model over-predicting by about 0.02 (14%) on average. However for sites of the TF CAD the real loss rate of 0.10 is grossly over-estimated by the model, at 0.24. We examine this anomaly in depth in the next subsection, and find it to point to self-cooperative DNA-binding by this TF.

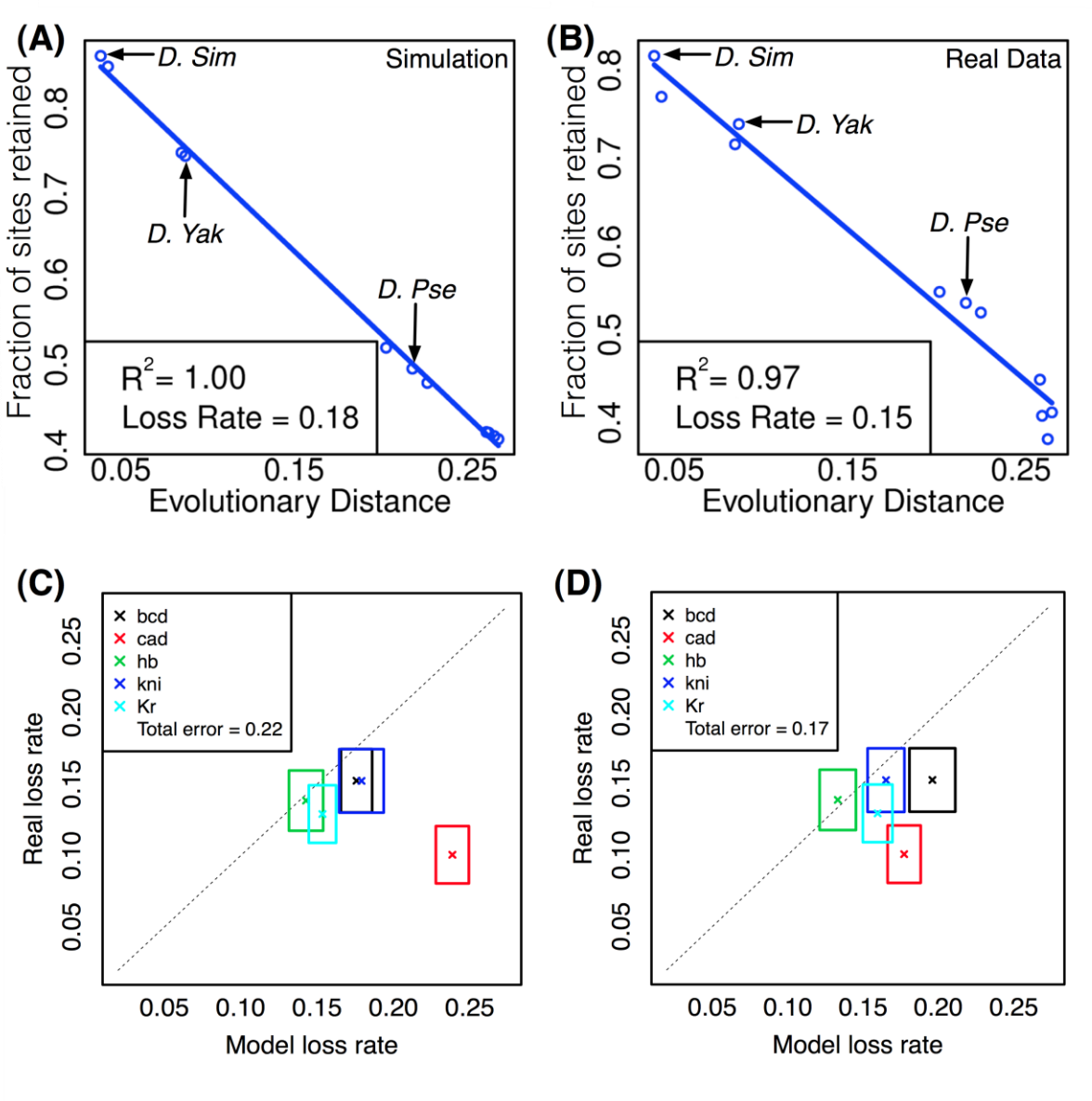


Figure 6.1: Modeling binding site loss rate. **(A-B)** Site conservation for BCD (y-axis) as a function of evolutionary distance between *D. melanogaster* and a second *Drosophila* species (x-axis), based on PEBCRES simulations (A) and real data (B). Evolutionary distance is measured as the average number of substitutions in aligned positions in the pairwise alignment (see methods). The inset shows the R^2 value and the (negative of) the slope of the best fit straight line, called the loss rate. **(C-D)** Site loss rate from real data (y-axis) and from PEBCRES simulations (x-axis), shown by cross marks for each TF. Sides of the each rectangle indicate the standard deviation of loss rates observed from bootstrap samples. The two panels show this information with two different models of regulatory function – one with self-cooperative DNA-binding by BCD and KNI (C) and one with self-cooperativity for BCD, KNI and CAD (D). The total error of a model was calculated as the horizontal distance between each cross and the diagonal, summed over all TFs, and is shown in the inset.

6.2.2 Evidence for CAD self-cooperativity

Figure 6.1 (C) reveals that the PEBCRES model shows reasonable agreement with observed site loss rates for all TFs except CAD. A similar disagreement was observed above (Figure 4.2) when comparing energy difference histograms of CAD sites from real data and simulations. As mentioned there, we hypothesized that this discrepancy may be due to self-cooperative DNA-binding by CAD. Such cooperativity has not been reported in the literature, and is not incorporated into the GEMSTAT model that was used in predicting genotype fitness values in our simulations. However, some evidence for such a mechanism was offered in the original analysis of Kim et al. (2009), where the distance between CAD sites was found to strongly correlate with loss rates, a potential signature of cooperative binding. A similar observation was made by Papatsenko et al. (2009). In the context of our analysis, such cooperativity may explain the apparent anomalies pertaining to CAD site evolution that are revealed by Figure 4.2 and Figure 6.1. If a pair of CAD sites act cooperatively, a model that ignores this effect will under-predict the fitness effect of a mutation in either site, and simulations based on such a model will lead to over-prediction of site loss.

Pursuing the above hypothesis, we modified the GEMSTAT model of CRM function to include CAD self-cooperativity, and retrained all model parameters on the 37 CRMs from *D. melanogaster*. We performed PEBCRES simulations again to predict the site loss rates for all TFs. Figure 6.1 (D) shows the results of this exercise, in the same format as Figure 6.1 (C). The new simulations predicted a loss rate of 0.17 for CAD sites, significantly closer to the real value of 0.10 than had been predicted above (0.24). The change in model affected predictions for other TFs but the overall agreement (see legend) for the model with CAD self-cooperativity was better than the model without it. We also repeated the experiments in Figure 4.2, now with the new model, and observed improved agreement with real data on CAD site conservation between *D. melanogaster* and *D. yakuba* (Supplementary Fig. S4 of (Duque et al. 2013)). We note that the GEMSTAT model in its default configuration (Figure 4.1, Figure 4.2 and Figure 6.1 (C)) incorporates self-cooperative DNA binding by BCD and KNI because He et al. (2010) found evidence for these mechanistic features by a statistical analysis of the same 37 *D. melanogaster*

CRMs that were studied by us. However, in that work, the evidence for CAD self-cooperativity was not statistically significant. In contrast, our analysis, which "fits" the GEMSTAT model to evolutionary data on those 37 CRMs via evolutionary simulations, suggests the presence of CAD self-cooperativity. Additionally, we repeated the above exercise with several alternative formulations of the GEMSTAT model, where we modeled self-cooperativity for the single TFs (BCD, CAD, HB, KNI and KR) and combinations of TFs (BCD and CAD; BCD and KNI; BCD, KNI and CAD) at a time. We found (Figure 6.2 B) that the evolutionary data on site loss rates is best explained by models that include self-cooperativity for CAD (e.g., a model that includes self-cooperativity for BCD, CAD and KNI, reported in Figure 6.1 (D)). These results can be viewed as evolutionary evidence for cooperative interaction between CAD binding sites. We also found that the spacing between neighboring CAD sites in *D. melanogaster* has a statistically significant bias for a range of 0-10 bp (base pairs), especially at 6 bp (Figure 6.3 A; see Methods), providing additional evidence for our hypothesis. (The sequence-to-expression model allows cooperative interactions between two homotypic bound sites that are within 50 bp of each other, and thus does not by itself suggest the preferred spacing between cooperatively bound sites.)

6.2.3 Experimental validation

We tested for direct physical interaction between CAD protein molecules using a variation of the LUMIER method (Barrios-Rodiles et al. 2005; Vizoso Pinto et al. 2009), modified to analyze direct binding *in vitro* (Cheng et al. 2013). A full length CAD coding region was fused to either luciferase (Luc) or maltose binding protein (MBP) and physical interaction was tested by measuring recovery of Luc-CAD following incubation with and purification of MBP-CAD. A sevenfold increase in recovered luciferase activity was observed with Luc-CAD compared with an unfused Luc control. This ratio is referred to as the Luminescence Intensity Ratio or LIR. In contrast, previously published negative control TF pairs all showed an LIR below 7 (Cheng et al. 2013; Kazemian et al. 2013). To further control for non-specific interactions, negative controls using unfused MBP, MBP fused to the CLK TF or Luc fused to CLK (Cheng et al. 2013; Kazemian et al. 2013) were also shown to result in lower recovery of luciferase. These results confirm the homo-dimerization of CAD molecules *in vitro* (Supplementary Table S1 of (Duque et al. 2013)).

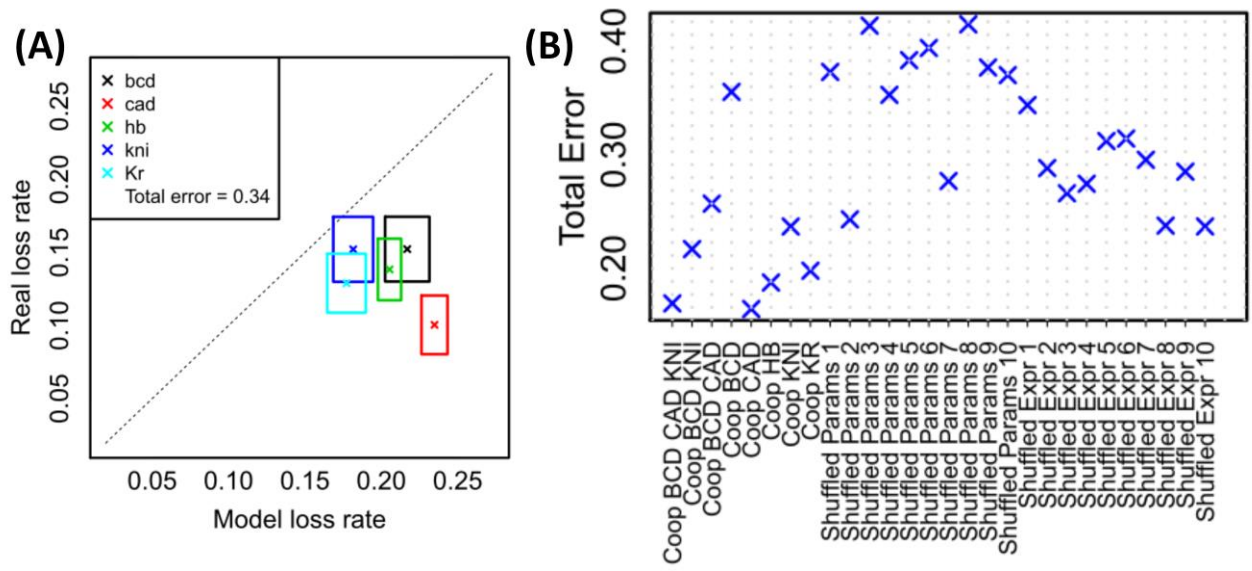


Figure 6.2: Negative Controls. (A) Real and simulation-based site loss rates (crosses) and their sampling variations (sides of rectangles), where simulations were performed with TF expression patterns randomly shuffled. **(B)** Total error, as defined above, for simulations performed with different configurations of the GEMSTAT model – self-cooperative DNA-binding by each of BCD, CAD and KNI ("Coop BCD CAD KNI", the model of Figure 6.1 D), by BCD and CAD only ("Coop BCD CAD"), by BCD and KNI only ("Coop BCD KNI", the model of Figure 6.1 C), by BCD only (Coop BCD), CAD only ("Coop CAD"), by HB only ("Coop HB"), by KNI only ("Coop KNI") and by KR only ("Coop KR") – and for different types of negative controls – with randomly reassigned TF parameters ("Shuffled Params 1-10") and with randomly reassigned TF expression profiles ("Shuffled Expr 1-10").

We next determined whether properly spaced pairs of CAD binding sites exhibited higher binding affinity than individual sites or the same sites with altered spacing. We identified two adjacent CAD binding sites with an optimal inter-site spacing of 6 bp (see Methods) and used a modification of a previously described oligo-binding assay (Hallikas and Taipale 2006; Cheng et al. 2013; Kazemian et al. 2013) by mixing luc-tagged TFs with biotin labeled DNA sites with an excess of unlabeled competitor DNAs. These competitors either match the wild type sequence or have mutations that alter the CAD binding sites or the spacing between them (Figure 6.3B). Differences in affinity are reflected in the ability of different competitor DNA molecules to prevent TF binding to the biotin-labeled DNA probe and thus reduce recovery of the associated luciferase activity with streptavidin beads. The wild type sequence containing both binding sites at the optimal spacing was the most effective competitor, reducing luciferase recovery to near background levels (Figure 6.3 C, Supplementary Table S2 of (Duque et al. 2013)). On the other

hand, when each site was provided on separate DNA molecules, or when both sites are on the same molecule but the spacing between the sites was increased by 5 bps, the competition was much less than with the wild type sequence, similar to the level seen with a single site. More detailed analysis revealed that an increase or decrease of 1 bp between the sites partly reduced binding while a change of 3 bp decreased binding to levels similar to that seen with a 5 bp change or a single site. From this result, we concluded that the CAD sites must be properly spaced for cooperative binding.

6.2.4 Negative controls

We claim above that the GEMSTAT model of CRM function, with self-cooperativity for BCD, KNI and CAD, provides the best fitness function to use with PEBCRES simulations in order to explain site loss profiles in the 12 *Drosophila* species. The total error (see Figure 6.1 legend) of loss rate predictions from this model is 0.17. We next performed two different types of negative control experiments where we did not expect the simulation-based loss rates to agree with data. These controls were intended to provide us a characterization of the total error values expected by chance. The effect of a TF on the expression of a CRM depends, among other things, on the thermodynamic parameters in the GEMSTAT model and the TF's concentration profile. In each set of controls, we randomized one of these factors while keeping the other factor unaltered. These are strong controls since most of the information contained in the original model is also present in the negative control.

In the first set of controls, we reassigned the thermodynamic parameters representing activation/repression strengths of TFs in the GEMSTAT model, in a random manner. For instance, if BCD (an activator) and KR (a repressor) have parameter values of +4 and -3 in the original model (positive and negative values signifying activation and repression respectively), the reassignment may assign a parameter value of -3 (repressive role) to BCD and a value of +4 (activating role) to KR. The reassignment is not necessary a simple swap between two TFs. For example, BCD might be assigned the parameters from KR, which receives KNI's parameters, while KNI is assigned the parameters from BCD. We performed 10 independent negative controls of this type, each with its own random reassignment of parameter values among TFs,

ran PEBCRES simulations of CRM evolution with the randomized GEMSTAT model, and recorded the total error of loss rate predictions. We found the best total error in these control experiments to be 0.24, with an average of 0.35 (Figure 6.2 B, “Shuffled Params 1-10”).

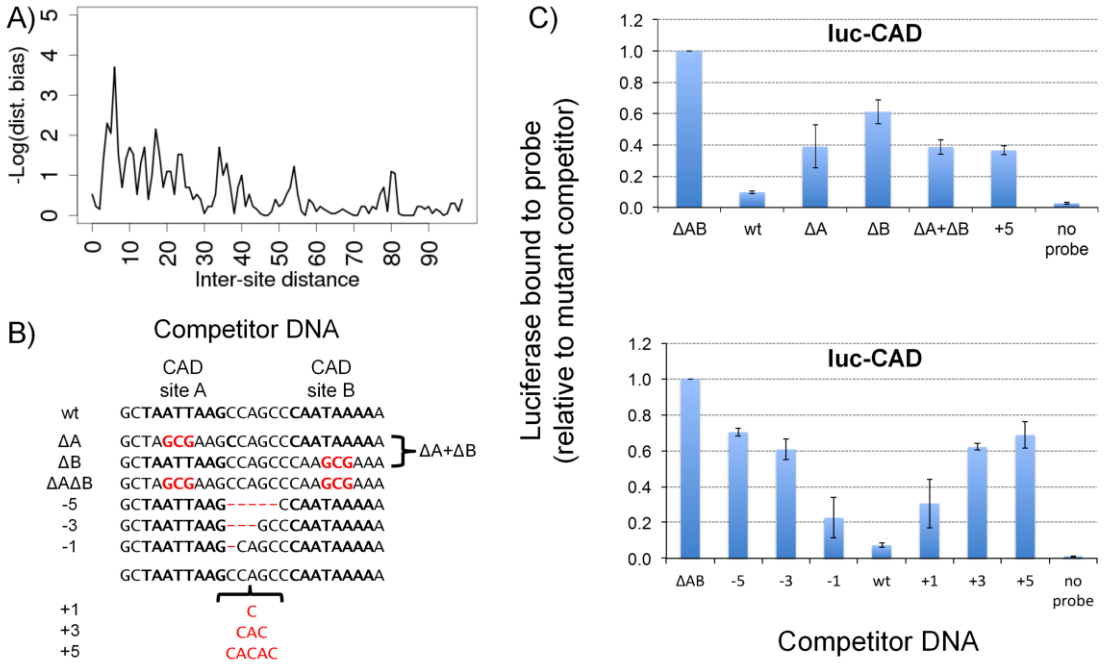


Figure 6.3: Experimental Validation. (A) Logarithm (base 10) of p-value of CAD inter-site spacing bias at different values of the spacing (x-axis). **(B)** A schematic representation of competitor DNA used to experimentally assess cooperative DNA binding by CAD *in vitro*. The competitor DNA might include mutations that disrupt one (ΔA, ΔB) or both (ΔAB) of the CAD-binding sites as well as deletions (-1, -3, -5) or insertions (+1, +3, +5) that change the spacing between the two sites. ΔA + ΔB indicates the inclusion of both DNA with mutations to the first site (ΔA) and DNA with mutations to the second site (ΔB). **(C)** DNA binding site measurements for CAD homotypic interaction. In experiments, the biotinylated DNA sequence is either wild type or not included ("no probe"). The competitor DNA used is indicated on the X-axis. The luciferase activity recovered using a competitor in which both CAD binding sites are mutated is set to a value of one and used as a non-specific DNA binding control to normalize the remaining samples. Addition of a wild type DNA sequences effectively competes for binding to the probe and reduces the recovery of Luc-CAD. Changes in either the individual CAD-binding sites or in the spacing between the binding sites results in reduced binding to the competitor DNA compared to wild type and an increased recovery of Luc-TF with the biotin-labeled DNA. Error bars indicate the standard deviation. See Supplementary Table S2 of (Duque et al. 2013) for individual measurements and more detailed sequence information.

In the second set of negative controls, we randomly shuffled the mapping between TFs and their expression profiles. For example, in the original model of anterior-posterior patterning in the embryo, BCD expression peaks in the anterior end and decays towards the middle of the embryo, while CAD expression peaks at the posterior end of the embryo and is weakest in the anterior end. A shuffled control might reassign these profiles so that BCD is active in the posterior end while CAD becomes active in the anterior. We repeated PEBCRES simulations ten times with this type of a randomized GEMSTAT model. Results from one such control are shown in detail in Figure 6.2 A, with a total error of 0.34. The best total error in these experiments is 0.24 and the average is 0.29 (Figure 6.2 B, "Shuffled Expr 1-10"). In summary, our negative control experiments confirm that the GEMSTAT model with self-cooperativity for BCD, KNI and CAD (total error = 0.17) provides an accurate explanation of site loss rates in 12 *Drosophila* species.

6.3 Summary and conclusions

In this Chapter we model the evolution of transcriptional binding sites across larger evolutionary timescales. We find patterns of site loss (Fig. 4) across these larger time-spans to be roughly consistent with predictions from a model that ignores positive selection, but there is much room for improvement in the goodness-of-fit. As such, we do not claim that site loss is adequately explained by purifying selection alone; in fact, some of the missing accuracy may be due to ignoring positive selection. PEBCRES simulations are not meant to be a test for positive selection, especially because the signal is mixed with the dominant signals of purifying selection acting on each CRM's output.

In trying to explain evolutionary data using our understanding of regulatory function, we also realized that the exact same framework may be used to test and improve our understanding of regulatory function using evolutionary data. We observed that the default configuration of the GEMSTAT model of regulatory function (used in the fitness function) led to evolutionary simulations that by and large agreed with real data on site evolution, but revealed one glaring disagreement – that for CAD sites. We took this as a cue that the GEMSTAT model of cis-regulatory logic may be flawed in some respect, and altered the model to include self-

cooperative DNA-binding by CAD. This led to much improved fits to evolutionary data, and subsequently the hypothesis of CAD self-cooperativity was experimentally confirmed both through PPI assays and DNA competition assays. Interestingly, two essential pairs of CAD binding sites have been previously described in the fushi-tarazu (FTZ) promoter (Dearolf et al. 1989), and a recent study (Bakkali 2011) of population variation in this promoter reported evidence of purifying selection, but the role of cooperative CAD binding and binding site spacing was not examined. Furthermore, one of the mammalian proteins related to CAD has been demonstrated to bind DNA as a dimer (Suh et al. 1994), indicating that dimer formation by members of this homeodomain family is conserved across species.

It is worth noting that a recent study by Kaplan et al. (2011) reported that protein interactions, including cooperative DNA-binding, play an insignificant role in determining TF occupancy at accessible regions of chromatin. We do not interpret their results as contradictory to our finding of self-cooperative DNA binding by CAD. The data type examined and modeled by Kaplan et al. is ChIP data on genome-wide TF-DNA binding levels, while we identified CAD self-cooperativity by modeling evolutionary data on CRMs and CAD binding sites within them. Moreover, our finding is not meant to be a broader statement on the prevalence of protein interactions in regulatory systems; it is only a demonstration of the possibility of hypothesizing such interactions through evolutionary analysis. The significance of this strategy for mechanistic investigation becomes clearer upon noting that the hypothesis of self-cooperative binding by CAD was also tested by He et al. (2010), in exactly the same expression-modeling framework (GEMSTAT) but on *D. melanogaster* CRMs alone, and not found to have significant support. It was only when we tried to explain CAD site evolution that an expression-model with CAD self-cooperativity appeared a much better alternative to a model without such cooperativity. We anticipate that there may be many more mechanistic insights about cis-regulatory logic that are not captured when we simply try to model expression from sequence, as in GEMSTAT, and will emerge only when we attempt to explain evolutionary data from such models. In this sense, our work may be a proof-of-concept of an entirely new strategy for modeling gene expression.

There are various technical issues involved in studying binding site evolutionary patterns that were addressed carefully by Kim et al. (2009), and we adopt their methodology throughout this work. One such issue is that of alignment errors. We performed all alignments using the PECAN program (Paten et al. 2009), which was shown by Kim et al. (2009) to lead to the same conclusions as those based on alignments from another program used there, called ProbConsMorph. A separate benchmarking study of alignment programs also found PECAN to be superior for aligning non-coding sequences (Kim and Sinha 2010). A second technical issue is that of binding site predictions, which, being based on motif matches alone, are prone to false positives. Again, this issue was addressed by Kim et al. (2009), who assessed the false positive rate for each of the TFs studied there. We excluded the TF *Giant (GT)* from our analysis as the estimated false positive prediction rate of its sites was high. In light of the same technical problem, we limited our study of site evolution on longer time scales to site loss events only, since gain events are more prone to being confounded with spuriously predicted sites.

We presented two closely-related evolutionary simulators, called PEBSES and PEBCRES, with the only difference being that PEBSES allows mutations only within a pre-designated binding site in the CRM and PEBCRES allows mutations anywhere in the CRM. While PEBCRES is a more realistic simulator, we do not dismiss the utility of PEBSES since (1) it was designed to match the SS and Halpern-Bruno models closely and therefore represents a fair comparison to these models, (2) it is computationally efficient for typical TFBS lengths (up to 20 bp long), and (3) it isolates the evolution of a site from the evolution of the nearby sites, allowing for the testing of different hypotheses.

The main caveats to note in this work are that both GEMSTAT and PEBCRES are imperfect models. There are aspects of gene expression, some known and perhaps several unknown, that are not encoded in the GEMSTAT model. The parameter learning procedure will, to a certain degree, compensate for mechanisms missing in the model by attributing their effects to other mechanisms. For instance, chromatin remodeling effects of pioneer factors (Harrison et al. 2011; Nien et al. 2011) that potentially make the local chromatin more accessible to other TFs may be inaccurately modeled as being distance-dependent cooperative binding between two

TFs. Likewise, there are many deficiencies in the evolutionary simulation framework adopted here, some of which are well known (e.g., not modeling several phenomena such as recombination, varying population size etc., and potential errors in evolutionary parameters used) but were not addressed by us for simplicity and efficiency. Additionally, our simulation framework relies on the assumption that the expression patterns do not change in any of the 12 *Drosophila* species. This is one of the reasons why we conducted this study on the segmentation network in the early *Drosophila* embryo, for which there is evidence of deep conservation at the gene expression level (Hare et al. 2008; Weirauch and Hughes 2010; Swanson et al. 2011). However, the assumption may not be valid for other systems of interest. Therefore, if evolutionary data does not agree with simulation results or agrees more with one model of regulatory function than another, one should treat this as merely suggestive of mechanistic hypotheses and as a starting point for further exploration.

In conclusion, we have presented here a new quantitative framework for exploring binding site evolution and cis-regulatory logic in an integrated manner. We show that this framework can offer a reasonable quantitative explanation of conservation and loss of individual TF-binding sites, and can also provide useful insights into biochemical mechanisms of gene regulation. This approach also has the potential to provide a theoretical framework for examining the outstanding issues of the day related to CRM architecture and evolution, such as homotypic clustering of binding sites (He et al. 2012), enhancer synergy (Yao et al. 2008), and shadow enhancers (Perry et al. 2010; Barolo 2012). Our future work will attempt to explain such phenomena using the general strategy presented here.

Future work: Detecting mechanistic features is fundamental task for better understating of regulation and evolution. New reports citing a variety of mechanistic hypotheses are constantly generated in the literature, with mechanisms including shadow enhancers (Perry et al. 2011; Barolo 2012), indirect activation (Kanodia et al. 2012), cooperative activation (Giniger and Ptashne 1988; Hertel et al. 1997), chromatin remodeling (Harrison et al. 2011; Nien et al. 2011), concentration specific roles for TFs (Papatsenko and Levine 2008), among many others. Many

of these hypotheses are generated based on very specific evidence, often times from a single biological system, and are often times hard to test experimentally.

It is therefore important to generate tools that can test these hypotheses against multiple available sources of data, be it from a single species or evolutionary data from multiple species. Our approach represents a promising and innovative way to generate and test hypotheses based on evolutionary data, as demonstrated in this section. However, our approach makes certain assumptions that while reasonable for our target system may not be appropriate in other contexts.

Particularly, our approach (a least in the form outlined in this section) assumes that the expression pattern does not change among the species being modeled. While this assumption is reasonable for the CRMs used in this section (Hare et al. 2008; Weirauch and Hughes 2010; Swanson et al. 2011), it may not be reasonable for other CRMs, especially those involved in latter stages of development or outside of developmental pathways. Moreover, it is often hard to obtain evolutionary data at the same resolution that is available for Drosophila. Also integral to our approach is the assumption that the trans-regulatory context (TF concentration patterns) and their biochemical properties, particularly their binding specificities, remain unchanged during the evolutionary span examined. These assumptions have received some support in the literature, e.g., (Fowlkes et al. 2011), but are not expected to be always true, even for early developmental systems and even within the *Drosophilids*.

Therefore we believe that an interesting focus for our future work is finding ways to detect mechanistic features that do not depend on the conservation of the expression pattern, do not assume the absence of adaptive selection and not do rely on abundant data for orthologous CRMs. This direction for future work can be explored in two ways: 1) by modeling and detecting mechanistic features without the use of evolutionary data or 2) by using evolutionary simulations to explore the fitness landscape in a manner similar to the one employed in Chapters 5 and 7.

The first approach is similar to that employed by He et al. (He et al. 2010) and involves using a thermodynamic model to fit experimental data and to check different models for their accuracy. This approach has been used with several models (Perkins et al. 2006; Korbel et al. 2007; He et al. 2010; Parker et al. 2011) and we are particularly interested in the approach by (Cheng et al. 2013) to detect interacting patterns by modeling ChIP data (ENCODE Project Consortium 2011).

The second approach involves the use of evolutionary simulation to the relative time necessary to evolve a specific expression pattern given the presence, or absence, of one or more mechanistic features. This approach is similar to our own approach described in Chapter 5 as well as the approach by Cooper et al. (Cooper et al. 2009). Cooper et al. (Cooper et al. 2009) study how likely a gene network is to produce a bimodal pattern given a set of features (e.g., the existence or not of competitive binding). This approach is also aligned with our objectives for Aim 4 and will be further discussed in Chapter 7. It is important to note that this approach does not necessarily need to match the results to real data. Therefore it does not make any assumptions regarding type of selection or the conservation of expression patterns and does not need a comprehensive library of orthologous sequences.

7 Understanding the evolution of regulatory sequences

In this chapter we describe a new type of biological insight we have learned using our tools. The questions we are interested in here are: "how long does it take for a CRM to evolve" and "what factors affect the evolution of a CRM?". This chapter is organized as follows: Section 7.1 describes the problem we are addressing and presents the motivation and related literature. Section 7.2 describes our results and insights. Finally, Section 7.2.1 summarizes our work and presents our main conclusions.

7.1 Background and motivation

The time scale for evolutionary changes in regulatory systems is an open problem that has puzzled biologists for over a decade, since it was first noticed that the general organization of certain regulatory sequences can be maintained for millions of years (Damjanovski et al. 1998) despite evidence that functional differences can evolve over significantly shorter time scales (Ross et al. 1994), and sequence comparisons showing that TF binding sites could appear and disappear among closely related species and even within a population (Damjanovski et al. 1998; Segal et al. 1999). This observation led Stone and Wray to approach the following question as a first step towards solving this puzzle: "*what time period would be required for new transcription factor binding sites to evolve (…) as a consequence of local point mutations (…) under the assumption of neutral evolution?*" (Stone and Wray 2001). They estimated that new binding sites can emerge due to point mutations alone in extremely short time scales (e.g. about 24 years in Drosophila or about 5,950 years in humans) even in the absence of selection. See (Stone and Wray 2001) for detailed numbers and assumptions.

The work by Stone and Wray opened a debate over the timescales necessary for the emergence of single binding sites. The main problem with their approach, as pointed out by MacArthur and Brookfield (2004) and Durrett and Schmidt (2007, 2008) is that their computation assumed independent evolution of every individual in the population, while in reality dependencies in the population are present due to common descent. Durrett and Schmidt (2007) addressed this issue, finding that the average time for a perfect 6bp binding site to appear in humans would be

nearer to 100,000 years and that the time for the appearance of a perfect 8bp binding site might be as high as 650 million years, implying that evolution of regulatory sequences is a very slow processes. However, if the requirement of a perfect match to the binding site is relaxed, as it is typically the case in nature, the time to evolve a "fuzzy" 8bp binding site would be approximately 60,000 years, which supports the possibility of changes to regulatory sequences on a short timescale.

The main issue with the approaches of Stone and Wray (2001) and Durrett and Schmidt (2007) is that they model the evolution of a single binding site, while in reality binding sites function and evolve in the context of CRMs, which in turn are composed of several binding sites for multiple TFs. This is an important issue since, as Stone and Wray (2001) point out, it is unlikely that the dozens of binding sites present in typical Drosophila CRMs could emerge one by one, under neutral selection, before other binding sites get destroyed. Therefore the models of Stone and Wray (2001) and Durrett and Schmidt (2007) have limited practical applicability, unless selection and the presence of multiple binding sites are properly accounted for.

To exemplify the issue with evolutionary models that operate at the level of individual binding sites, we can look at compensatory mutations. This is the phenomenon where a deleterious mutation is followed by a mutation elsewhere, typically in the same CRM, that compensates for the deleterious effect of the first mutation. Initially, the prevalent theory was that the deleterious mutation would fix due to genetic drift and, subsequently, the compensatory mutation would fix due to positive selection. This theory predicts such turnover events to be more common on smaller populations (such as vertebrates compared to invertebrates), since drift is accentuated. However, sequence comparison data shows that turnover events are in fact more common in larger invertebrate populations, where the effect of genetic drift is expected to be more limited.

Such a paradox could not be explained by modeling the evolution of binding sites as independent events, and was notably addressed by Carter and Wagner (2002). By explicitly modeling the evolution of pairs of binding sites, instead of single binding sites, they identified

the cause of this paradox to be the phenomenon called *stochastic tunneling* (Iwasa et al. 2004), in which the second (compensatory) mutation happens before the first (deleterious) mutation is fixed. The result of including *stochastic tunneling* in the population genetics model is that turnover events become more common for populations of size similar to invertebrate populations (Carter and Wagner 2002).

Durrett and Schmidt (2008) expanded on the work of Carter and Wagner (2002) in an attempt to obtain practical estimates of the time necessary for a mutation in a binding site to happen following a mutation in another binding site. They claimed that in Drosophila a pair of mutations can inactivate a binding site and activate another on the timescale of several million years, consistent with observed results that point to rapid turnover of binding sites in Drosophila.

MacArthur and Brookfield (2004) approached the question regarding the timescale necessary for the evolution of regulatory sequence from a different perspective; instead of focusing on the time necessary to for new binding sites to emerge, MacArthur and Brookfield (2004) were concerned with the time necessary for an entire CRM to emerge. They observed that the time to evolve a CRM that drives a certain level of activation by a TF depended on the CG-content of the sequence that served as starting point for the evolutionary simulation. They also noted how distance-dependent cooperativity influences which "pre-sites" in the starting sequence evolve into functional sites. However, their model falls short in two aspects: First, it did not explain the emergence of real enhancers since they did not model the combinatorial regulation present in real enhancers, and instead focused on simple enhancers responding to a single activator TF (in a manner similar to our work presented on Chapter 5). Second, their work does not provide any real time estimate for the emergence of a CRM, possibly due to an excess of simplifying assumptions that had to be made in their model.

While all of these methods have provided pieces to the puzzle of how much evolutionary time is necessary to evolve a CRM, there does not exist, to the best of our knowledge, a computational model capable of estimating the time necessary to evolve a complex CRM

involving multiple TFs with distinct roles, under a range of population genetics and mechanistic assumptions. In this chapter we will demonstrate how PEBCRES can be used for exactly this purpose, and share some of the biological insights that can be learned from such exercise.

7.2 Results

7.2.1 Overview of simulations

We used the PEBCRES simulation framework (He et al. 2012; Duque et al. 2013) (Figure 7.1 A,B) to evolve sequences that drive a pre-determined expression pattern, simulating the process of evolutionary adaptation under a variety of scenarios. The main simplifying features of a PEBCRES simulation are: (1) a constant sized population of 2N haploid individuals evolves as per the Wright-Fisher model (Wright 1931; Fisher 1999), (2) each individual's genotype is a DNA sequence 500 – 2000 bp long (typical length of a CRM), (3) mutations occur at a fixed rate and independently at each nucleotide, and (4) no recombination occurs. Selection is modeled so that an individual *i* spawns an expected number of offspring proportional to $1+KF_i$ where K is a constant called the "selection scale" and F_i is the fitness of individual *i* on a scale of 0 (unfit) to 1 (fit). Additional details in (He et al. 2012; Duque et al. 2013).

The distinguishing feature of a PEBCRES simulation is its calculation of a fitness value (*F*) for any given CRM-length sequence and a given expression pattern called the "target pattern". The target pattern is pre-specified as a (say M-dimensional) vector of gene expression values on a scale of 0 to 1 (Figure 7.1 A). The sequence is mapped to the expression pattern it encodes (also an M-dimensional vector) by the statistical thermodynamics-based GEMSTAT model (He et al. 2010). Note that the parameters of GEMSTAT, representing the trans context, are trained before and outside of PEBCRES simulations. (Also see next paragraph for comments about reliability of these parameters.) The predicted expression pattern corresponding to the sequence is then compared to the target pattern by a specialized function called "weighted Pattern Generating Potential" or wPGP (Duque et al. 2013; Samee and Sinha 2013) to produce a fitness value between 0 and 1, which is 1 if and only if the two pattern vectors are identical.

To decide on the target expression patterns to use in our study, we considered a set of 37 *bona fide* CRMs from *D. melanogaster* that drive well characterized anterior-posterior (A/P) patterns in the blastoderm stage embryo. These 37 CRMs were the subject of a detailed modeling exercise in our previous work (He et al. 2010), and the accuracy of model fits for a majority (see Supplementary Note A.1 on Appendix A) of CRMs in that exercise assure us that the genotype-to-phenotype mapping used in PEBCRES simulations here is a reasonable approximation of reality. Furthermore, in (Duque et al. 2013) we analyzed the evolutionary changes within these 37 CRMs across the *Drosophila* sub-family (12 sequenced species separated by ≤ 65 Myrs) and were able to accurately model these changes using PEBCRES simulations of a functionally constrained CRM. The selection strength on CRMs (i.e., the selection scale parameter K mentioned above) estimated in that study as providing the best fits between model and data was used as the default value in the current study. We selected 28 of the 37 A/P patterning CRMs as the subject of our analyses (see Methods for selection criterion), predicted their expression patterns using GEMSTAT, and used these 28 predicted patterns (see Supplementary Figure A.1 on Appendix A), which are in approximate agreement with experimental CRM readouts, as the target patterns in PEBCRES simulations. We will refer to each target expression pattern by the name of the *D. melanogaster* CRM associated with that pattern.

Thus, using a carefully constructed fitness function and with target patterns representing the typical complexity of a developmental CRM, we hoped that our simulations will provide meaningful insights into what it takes to evolve an enhancer.

7.2.2 Estimating the time to evolve a CRM

Our first goal was to estimate how long it might take for a typical developmental CRM to evolve from genomic background, under a variety of assumptions. We simulated the evolution of random sequences targeting each of the 28 target patterns (at least 30 simulations for each pattern) and recorded the "time-to-evolve" for each simulation, i.e., the earliest generation in which an individual with fitness above 0.8 emerged in the population. (We noted that fixation quickly follows the emergence of a fit genotype.) Since our simulations are constrained by limited computational resources, we imposed a maximum number of generations (100,000) on

all simulations (see Supplementary Note A.2 on Appendix). Using ideas from population genetics theory and properly accounting for the time rescaling used by PEBCRES (see Section 3.4.2) (Hoggart et al. 2007; He et al. 2012), we converted the time-to-evolve value from generations to an estimate of time in millions of years of *Drosophila* evolution. Finally, we examined the median over all of our simulations for each target pattern. The results of this computational experiment are presented in Figure 7.1 C, and discussed below.

Our simulations predict that, under strong selection, functional CRMs for complex spatial patterns could evolve in surprisingly short evolutionary times. For example, the average time necessary to evolve the pattern for gt_-10, as per our simulations, is only ~0.3 million years. As a point of reference, this is nearly 10 times smaller than the divergence between *D. melanogaster* and *D. simulans* (2.5 million years (Ranz et al. 2003), synonymous substitution rate of ~0.04 (Bedford and Hartl 2008)), predicting that even between these two closely related species there should be lineage-specific CRMs driving simple expression patterns defined by the response to a single TF. (The 'gt_-10' pattern is mediated by activating sites of the Bicoid (BCD) transcription factor.) Other quickly evolving patterns were mostly BCD-driven anterior patterns like 'gt_-10', but also included more central patterns such as 'h_stripe_34_rev' and 'run_stripe5' (Figure 7.1 C), which are regulated by two or more TFs (Supplementary Figure A.1 on Apendix A).

On the other hand, some target patterns require much longer time to evolve, with the longest time being about 9 Myrs (median) for the expression pattern 'kni_83_ru', roughly 30 times longer than that for 'gt_-10'. There is a clear trend of anterior patterns to have lower time-to-evolve estimates while central and posterior patterns have larger estimates (Figure 7.1 C, bottom). We noticed that half of the expression patterns have time-to-evolve estimates that are higher than the distance between *D. melanogaster* and *D. simulans* (2.5 million years (Ranz et al. 2003), the closest of the currently sequenced species), while all of the patterns have time-to-evolve that is shorter than the divergence between *D. melanogaster* and *D. yakuba* (13-17 million years (Satta et al. 1987; Satta and Takahata 1990)). This suggests an opportunity for future studies to compare these sequenced genomes, which are amenable to high quality

alignments, for the existence and function of many lineage-specific CRMs. Our theoretical findings are also supported by the recent discovery of hundreds of CRMs (driving expression in *Drosophila* S2 cells) being gained since the *D. melanogaster – D. yakuba* split [CITE doi:10.1038/ng.3009].

7.2.3 Evolutionary sampling of the fitness landscape: real vs *in silico* evolved CRMs

We next examined the *in silico* evolved CRMs (also called 'simulated' CRMs below) from the previous section more closely, with a view to gain deeper insights into the 'fitness landscape' (Berg et al. 2004a) associated with each target expression pattern. Our primary goal was to determine (1) if these simulated CRMs resemble the real *D. melanogaster* CRM associated with the target pattern, as might be expected, and (2) whether cases that deviate from this expectation provide clues about shortcomings in our models of CRM function (Duque et al. 2013), reveal signatures of the evolutionary process (He et al. 2012) or suggest multiple optima in the fitness landscape. For this investigation we chose to describe a CRM by the estimated 'occupancy' of each TF in the CRM (see Methods (He et al. 2012)), which is an integrated score reflecting the total number of binding sites, both strong and weak, of that TF (Also see Supplementary Note A.3 on Appendix A) It also enables easy comparison of two CRMs for similarity of cis-regulatory logic. We compared any two CRMs, real or evolved, by the Euclidian distance between their respective six-dimensional vectors of TF occupancy counts (GEMSTAT modeling was based on six TFs).

We first examined all *in silico* evolved CRMs for all 28 target patterns and noted that CRMs associated with similar expression patterns are closer to each other than distinctly expressed CRMs (Supplementary Figure A.2 on Apendix A), as expected. We then asked if *in silico* evolved CRMs for the same target pattern cluster in the vector space, and how tight these clusters are. Table 7.1 presents two relevant metrics to answer these questions. The first metric, d_{intra}, represents the average distance between any pair of evolved CRM for a particular target pattern, and a second metric, $d_{inter,}$ denotes the average distance between CRMs for a specific expression pattern and CRMs representing other patterns. (We restricted the other patterns to be those that are least correlated with that pattern, since several of the target patterns are

highly similar to each other.) As Table 7.1 shows, the ratio d_{inter}/d_{intra} is almost always ≥ 2, indicating that distinct target patterns are associated with well-clustered simulated CRMs. A few examples are depicted in Figure 7.2 (note black circles in each panel), which further confirms this observation.

We next asked if *in silico* evolved CRMs for a target pattern are similar to the *D. melanogaster* CRM (henceforth, 'real' CRM) associated with that pattern. For this, we calculated a metric, d_{WT}, as the average distance between the real CRM and all simulated CRMs for each target pattern, and compared it to the inter-cluster distances d_{inter} as well as intra-cluster distances d_{intra} defined above. A large relative value of d_{WT} indicates that CRMs resulting from evolutionary simulation are different from the real CRM. As Table 7.1 shows, d_{WT} tends to be slightly larger than d_{intra} but smaller than d_{inter}, indicating that the real CRM falls more or less within the cluster of evolved CRMs for the same expression pattern (Figure 7.2 A,B,C,D). There were a few interesting exceptions to this trend, marked with a † superscript in the table. For example, the *in silico* evolved CRMs for kni_83_ru and h_15_ru (Figure 7.2 E,F) seem to be distinctly more parsimonious than the real CRM, although GEMSTAT predicts their functionality to be the same (also see Supplementary Note A.4 on Appendix A). We may speculate on why high occupancy evolved in the real CRM for these patterns. One hypothesis is that the evolutionary history of the real CRMs is more complicated than our simple simulations assume, e.g., they have been 'exapted' (de Souza et al. (2013)) from other functional sequences to perform a different function. An alternative possibility is that the high occupancy values seen in the real CRM are functionally necessary due to some unknown mechanism not modeled by GEMSTAT.

7.2.4 Features of CRM composition may influence its time-to-evolve

As noted above, time-to-evolve estimates for CRMs of different expression patterns vary greatly, by at least one order of magnitude. We sought to determine the factors that can explain such variability, focusing on two classes of potential determinants: binding site content of the CRM, and features of the target expression pattern itself.

We first tested for a correlation between time-to-evolve for a target pattern and each TF's binding site count (or estimated occupancy) in the real CRM associated with that pattern. We found that binding site content of the TF HB has a strong positive correlation with time-to-evolve estimates (Pearson CC = 0.70, p-value = 1.5×10^{-5}, Figure 7.3 A). We also found that total binding site content of a CRM, aggregated over all six TFs, significantly positively correlates with time-to-evolve estimates (Figure 7.3 B); however, this effect can be attributed mostly to HB site content, as indicated by a weak partial correlation coefficient (Johnson et al. 1992; Whittaker 2009) with p-value of 0.45.

We next asked if certain aspects of the target pattern make it harder to evolve. A visual inspection (Figure 7.1 C) suggested that expression in the anterior domain of the embryo marks smaller time-to-evolve estimates. To probe this point further, we calculated the Spearman's Correlation Coefficient between the expression level of a CRM at a fixed position along the A/P axis and the time-to-evolve estimate of that CRM, and repeated this procedure for every axial position. We found strong negative correlation at anterior positions (Figure 7.3 C), i.e., anterior expression patterns appear to be easier to evolve. We noted also that the plot of correlation coefficients in Figure 7.3 C very closely a resembles 'flipped' version of the expression pattern of HB (Figure 7.3C, dashed line), suggesting again that the faster evolution of CRMs with anterior patterns may be related to their HB binding levels. This is consistent with the fact that HB is modeled in GEMSTAT as a repressor, and therefore high levels of expression in the anterior end of the embryo indicate absence of HB sites in the CRM, which in turn correlates with shorter time-to-evolve estimates. We find it surprising that a single TF correlates so strongly with time-to-evolve estimates, and speculate that it may be due to the repeat-like T-rich motif of Hb (Supplementary Figure A.3 on Apendix A), or an artifact of mechanistic details about HB regulation not captured in GEMSTAT (see Discussion).

Finally, we find that the number of TFs involved in generating the pattern also correlates with high time-to-evolve estimates for that pattern (Figure 7.3 D), with Pearson's correlation coefficient of 0.49 (p-value of 0.005). We calculated the number of TFs needed to generate a

pattern as the number of TFs that have at least one site above the LR threshold of 0.25, but the correlation remains significant for other thresholds on the strength of sites (data not shown).

7.2.5 Dependence on initial conditions, and the possibility of exaptation

Recall that each of our simulations begins with a random sequence. If the initial random sequences have a higher fitness value for certain target patterns, perhaps due to a greater frequency of random occurrence of certain binding sites necessary for that pattern, then such patterns may be quicker to evolve. This is the reason why we selected only 28 expression patterns out of the 37 A/P expression patterns modeled in (Duque et al. 2013) (see Methods). Even within these 28 target patterns, we observed a significant positive correlation between the average fitness of random (initial) sequences and median time-to-evolve estimate (Supplementary Figure A.4 on Apendix A). However, a partial correlation analysis (Johnson et al. 1992; Whittaker 2009) revealed that this correlation with fitness of initial sequences is not significant if we discount the already noted correlation with HB site counts in the real CRM. This was not true when partialing out the effect of other TFs' sites counts (data not shown). Moreover, the correlation between HB site content and estimated time-to-evolve remains significant after partialing out the effect of initial fitness (data not shown). We interpret these observations to suggest that of the number of HB sites in the initial random sequences influences the fitness of those sequences for certain target patterns, and therefore their time-to-evolve estimates.

Simulations beginning with random sequence represent an extreme scenario of evolution of regulatory sequences. In reality, features of the initial sequence where a CRM is to arise may strongly influence the waiting time. For instance, as previously noted (MacArthur and Brookfield 2004; Durrett and Schmidt 2007, 2008) , the composition of the genomic background affects the time required to evolve binding sites and regulatory sequences. Dermitzakis et al. (2003) noted that CRMs have short words that are close to becoming functional sites, and thus have the potential to quickly gain new function. Taking this line of reasoning further, one might argue that a CRM may readily evolve by transformation of a sequence that already contains several relevant binding sites (Prud'homme et al. 2007; Okada et al. 2010; Emera et al. 2012; de

Souza et al. 2013), a scenario that may be considered as an example of exaptation, also known as co-opted evolution (Hoekstra 2006).

We designed two computational experiments to explore the effect of initial sequences on time-to-evolve. The first experiment simulates evolution under the favorable scenario where a CRM evolves from a sequence that drives an expression pattern very similar to the target pattern (Methods). The second experiment explores an opposite scenario, in which a CRM evolves from a sequence that drives a very different pattern (for example, in which a CRM with anterior expression evolves from a sequence that drives posterior expression). As expected, the time to evolve each of the CRMs in the first experiment is largely reduced (Figure 7.4 A) due to the abundance of binding sites for the necessary TFs. However, simulations from initial sequences that drive a pattern anti-correlated with the target has a negative effect on evolutionary time of several CRMs (Figure 7.4 B). This is due to the contrasting roles that some pairs of CRMs have. Binding sites present in the initial sequence are expected to reduce time-to-evolve only if they are for the right TFs, i.e., ones that can contribute to the target pattern. If, on the other hand, the starting sequence has several sites that disrupt the target pattern and few sites that contribute to it, evolution will have to proceed by deconstructing the initial sequence before it can start constructing the target pattern. For example, the 'kni_83_ru' CRM drives expression in a stripe in the posterior end of the embryo (Supplementary Figure A.1 on Apendix A), and contains many binding sites for CAD, GT, HB and KR. If we used this sequence to initiate simulations for the target pattern 'eve_1_ru', a stripe in the anterior end of the embryo, the evolving sequence would have to lose most of its binding sites for CAD and HB, maintaining the sites for KR and gaining new sites for BCD.

7.2.6 Uniformly expressed activators can speed up emergence of CRMs

Patterning of the early *Drosophila* embryo is well known to be achieved by gradients of maternally deposited transcription factors and by their patterned regulatory targets. Recent studies have focused also on uniformly expressed TFs that function as in important activators in patterning systems (Liang et al. 2008; Harrison et al. 2011; Tsurumi et al. 2011). These activators by themselves do not or may not have the patterning ability of non-uniformly

expressed TFs, but can modulate the response of a CRM to a patterned signal (Kanodia et al. 2012). They are present in several regulatory systems including the A/P system (Arbouzova and Zeidler 2006; Liang et al. 2008; Kanodia et al. 2012), the Dorsal-Ventral (D/V) patterning system (Liang et al. 2008; Kanodia et al. 2012) and other patterning or developmental systems (Arbouzova and Zeidler 2006; Nien et al. 2011; Tsurumi et al. 2011). Here, we pursued the hypothesis that the deployment of uniformly expressed activators in patterning systems also has an evolutionary explanation: that they improve the "evolvability" of target patterns, by increasing the number of viable paths evolution can take from a random initial sequence to a functional CRM.

To explore this hypothesis we repeated the time-to-evolve simulations from above with a GEMSTAT (CRM function) model specification that includes a ubiquitous activator, and compared the results to those from the original model. We designed a methodology that ensures that there exists a fit solution (CRM) for the target pattern under either function model, with and without the ubiquitous activator, so that any difference in time-to-evolve can be attributed to the evolutionary ramifications of the ubiquitous activator (see Methods). We tested the effects of two well-characterized ubiquitous activators, ZLD (Liang et al. 2008; Harrison et al. 2011) and DSTAT (Tsurumi et al. 2011), separately. As shown in Figure 7.5, each of these TFs reduces the median time-to-evolve for several target patterns, with the effect of ZLD being clearly more prominent. A Two-Way Analysis of Variance supported these observations, with P-value of 2×10^{-4} (ZLD) and 0.03 (DSTAT) (Supplementary Tables A.1, A.2 in Apendix A), indicating that adding either ubiquitous activator to the model has a statistically significant effect of decreasing time-to-evolve.

For deeper insights into the effect of ubiquitous activators on time-to-evolve, we discuss the example of the target pattern 'eve_37ext_ru', which comprises a single stripe of expression peaking at about 49% egg-length (Figure 7.6 B). (This is the third stripe of *eve* expression along the A/P axis, with stripe 7 being outside the modeled range of 20-80% egg length.) To drive this pattern using the TFs in the baseline model (BCD, CAD, GT, HB, KNI and KR), whose A/P expression profiles are shown in Figure 7.6 A, evolution could add activator sites for TFs BCD

and CAD, generating expression across all of the AP axis, and add repressor sites for KNI and HB to create repression at the anterior and posterior sides of the desired stripe. Indeed, this may be the strategy employed by nature, since these are the TFs for which sites are present in the real CRM from *D. melanogaster* (Figure 7.6 C). However, neither BCD nor CAD has maximal concentration around 49% egg-length (Figure 7.6 A), and to create sufficient activation in the central domain of the A/P axis it would be necessary to add several strong sites to the CRM. On the other hand, if DSTAT is also available (as a ubiquitous activator), evolution could use DSTAT sites to add to the weaker activation by BCD and CAD in the central domain. This offers another avenue for evolution to explore, ultimately leading to a lower time-to-evolve in our simulations. Intriguingly, the *D. melanogaster* CRM for 'eve_37ext_ru' has two DSTAT sites (not shown), suggesting that this may indeed have been the avenue taken by evolution. Our interpretation is in agreement with theories of evolutionary computation (Holland 1975; Goldberg 1989, 2002), according to which if a combinatorial problem has many fit solutions we are more likely to find one of these solutions quickly (Goldberg 2002).

7.2.7 Sensitivity to evolutionary parameters

We began this study by estimating the time necessary to evolve 28 different expression patterns starting from a random sequence. These estimates are expected to depend on values of the population genetics parameters used in the simulations, in particular the population size N, the mutation rate μ and the selection coefficient s. We explored these dependencies next, varying the simulation parameters within reasonable ranges.

All of our simulations used a time rescaling heuristic (Hoggart et al. 2007; He et al. 2012) for speeding up simulations, with scaling factor λ=1000, a time-scaled population size $2N$ = 1000 and a time-scaled mutation rate $\mu = 10^{-5}$ (mutations per generation per base pair), resulting in a scaled mutation rate $2N\mu = 10^{-2}$, which is within the estimated range of $10^{-2} - 10^{-4}$ (Drake et al. 1998; Thornton and Andolfatto 2006) for Drosophila (see Section 3.4.2). We note however that this mutation rate is higher than that used in (Duque et al. 2013). The higher mutation rate reduces the computational time required for a simulation and as mentioned, is still within the estimated range for Drosophila. However, to understand the effect of mutation rate on our

results, we repeated the time-to-evolve estimation procedure with values of $2N\mu$ that are an order of magnitude greater or lesser than 10^{-2}. Figure 7.7 A shows how the time to evolve a CRM, averaged over the 28 target patterns, changes with the values of $2N\mu$. Changing the scaled mutation rate $2N\mu$ by a factor of 10 results in time-to-evolve estimates that change by less than 10 times, which is not unexpected since different values of $2N\mu$ result in different balances between selection and drift. In particular, reducing $2N\mu$ from 0.01 to 0.001 (a factor of 10) results in average time-to-evolve increasing about 7-fold from ~2.1 Myrs to ~18 Myrs, with estimates for individual target patterns ranging between 2.1 Myrs and 25 Myrs. (As a comparison point, we note the estimated divergence time between *D. melanogaster* and *D. pseudoobscura* to be 25-55 Myrs (Richards et al. 2005).)

Another important population genetics parameter is the selection coefficient *s*, or equivalently, the population-scaled selection coefficient *4Ns*. In our simulations the strength of selection is controlled by the selection scale parameter *K*, which is analogous to *s* when two competing individuals have fitness of 0 and 1. For the experiments reported above we used $K = 50$, which is of the same order as the value determined in (Duque et al. 2013) to provide the best fit to real evolutionary data. In the absence of better tools to estimate the actual strength of selection, this value is our best guess in the context of our experiment. Nevertheless, we repeated our experiments with different values of *K* (5, 25, 50, 100), as shown in Figure 7.7 B, in part to compensate for our lack of knowledge of the real selection strength and in part to understand how the selection strength influences the time-to-evolve. As expected, smaller values of the selection scale *K* result in longer times necessary to evolve CRMs. For instance, reducing the selection scale by a factor of 10 ($K = 5$) results in time-to-evolve estimates increasing by less than a factor of 10. We also found an apparent saturation in the effect of increasing the selection strength from 50 to 100 (Figure 7.7 B).

Finally, we note that other assumptions about the evolutionary model might also influence time-to-evolve estimates. For example, insertions and deletions (indels) have been suggested to have important effects on the evolution of regulatory sequences (Sinha and Siggia 2005; Lusk and Eisen 2010; Nourmohammad and Lässig 2011); recombination has also been suggested to

influence the rate of adaptation (Schoustra et al. 2007) and even ploidy had been suggested to influence adaptation (e.g. (Orr and Otto 1994; Zeyl et al. 2003)). It is beyond the scope of this work to test for the effect of all such mechanisms, but we examined the effects of indels on time-to-evolve estimates. We find that adding indels (insertions implemented as short tandem repeats, as in [CITE]) to our model increases time-to-fit estimates from ~2.8 Myrs to ~3.6 Myrs on average, a statistically significant increase (Paired T-test with pooled standard deviation, P-value 0.0002). One way to interpret this is that insertions and deletions are more likely to completely destroy binding sites than point mutations, and therefore are more likely to be selected against.

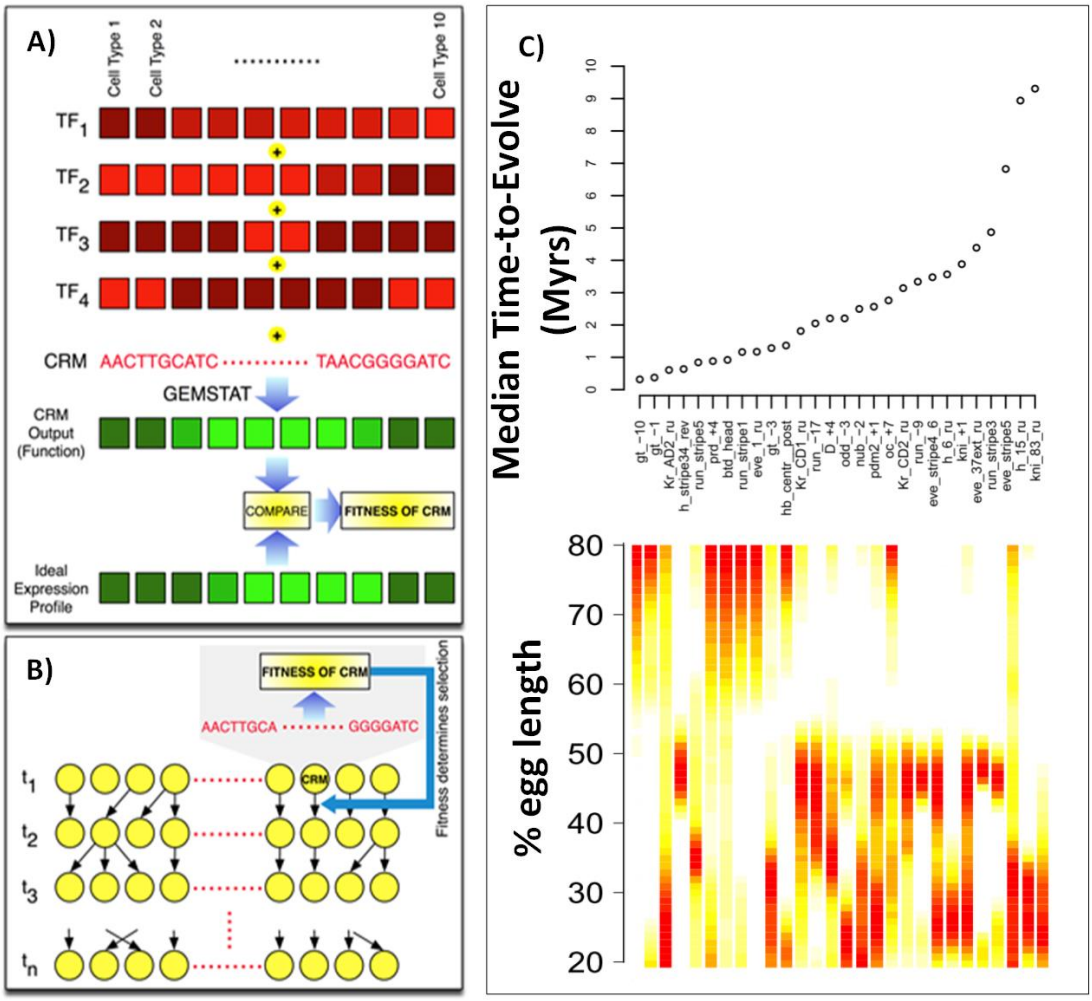


Figure 7.1: Estimating the necessary for a CRM to evolve. **(A,B)** Methodology**.** A schematic representation of the PEBCRES framework describing how it is used to estimate the time necessary for a CRM to evolve from genomic background. Expression readout of the evolving CRM is predicted using GEMSTAT, producing a fitness value (A), which is then plugged into a Wright Fisher Simulation with selection (B). **(C)** Top panel: time-to-evolve estimates (y-axis), in million years, for each of 28 target expression patterns (x-axis). Bottom Panel: A representation of the 28 A/P expression patterns that serve as target patterns in our simulations, sorted by time-to-evolve estimate (same order as in top panel). Each expression pattern is represented by a column in the heatmap, with red representing high expression and white representing absent expression. The anterior end of the embryo is at the top and posterior end at the bottom. Only 20-80% egg length interval is shown.

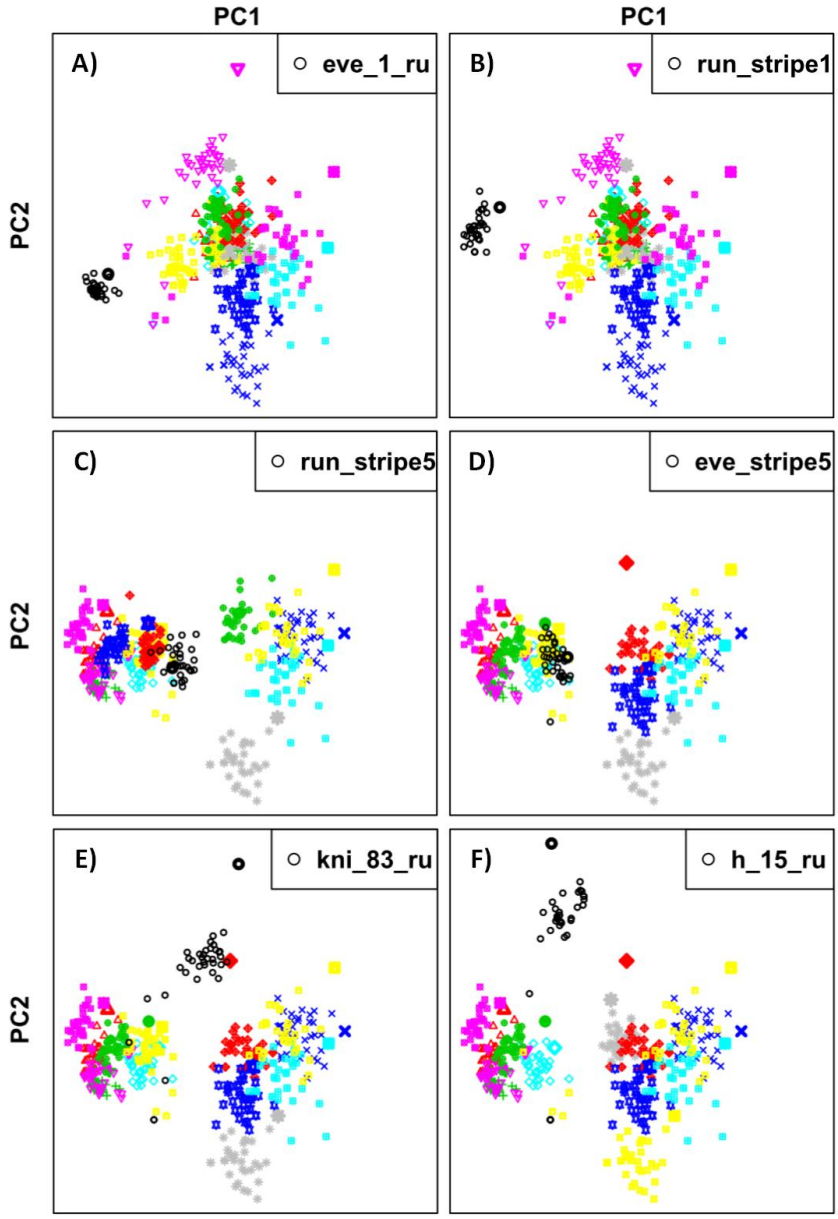


Figure 7.2: Visual representation of real and in silico evolved CRMs. The representation is given in a 2-dimensional projection of the 6-dimensional 'TF occupancy' space occupied by these CRMs. The axes represent the first and second principal components. The panels correspond to CRMs for patterns 'eve_1_ru' (A), 'run_stripe1' (B), 'run_stripe5' (C), 'eve_stripe5' (D) 'kni_83_ru' (E) and 'h_15_ru' (F). In each panel, simulated CRMs of respective pattern are shown in small black circles, and the real *D. melanogaster* CRM for that pattern as a larger black circle; points in other colors represent simulated CRMs (smaller icons) and the real CRM (larger icon, same color) for other target patterns.

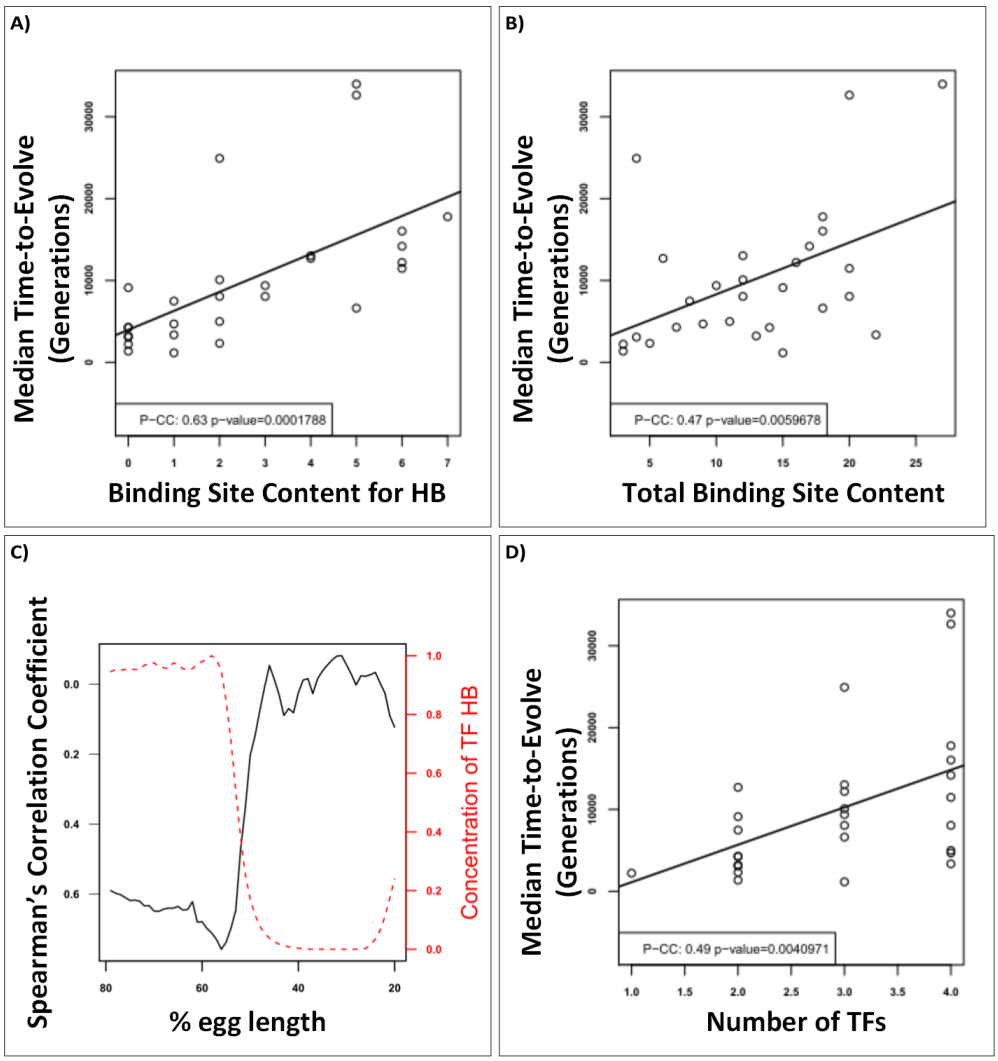


Figure 7.3: Features of CRMs that influence time-to-evolve. **(A)** A scatter plot relating the estimated occupancy of HB in a real CRM (x-axis) and the median estimated time to evolve a CRM for the corresponding pattern (y-axis). The Pearson Correlation Coefficient between the two variables is of 0.7, which is significant at a P-value of 1.4×10^{-4}. The best fit line is also shown (solid line). **(B)** A scatter plot relating the estimated TF occupancy in a CRM, summed over all TFs used in the model (x-axis), and the median estimated time-to-evolve for the corresponding pattern (y-axis). Pearson CC = 0.47, P-value = 0.005. However the partial correlation, discounting the contribution of HB sites, is not significant (P-value = 0.45). The best fit line is also shown (solid line). **(C)** Time-to-evolve estimates of CRMs are highly negatively correlated with expression level in anterior parts of the embryo. The y-axis shows for each position along the A/P axis ('%egg length', x-axis) the Spearman's Correlation Coefficient between a target pattern's expression level at that axial position and the time-to-evolve estimate for that pattern. The concentration profile of HB across the axis is also shown (dashed line). **(D)** Scatter plot relating the number of TFs with at least one binding site present in the *D. melanogaster* CRM (x-axis) and the median estimated time-to-evolve for the corresponding pattern (y-axis). The Pearson Correlation Coefficient between the two variables is 0.49, P-value = 0.005, indicating the number TFs acting in a pattern correlates with the Time-to-evolve that pattern.

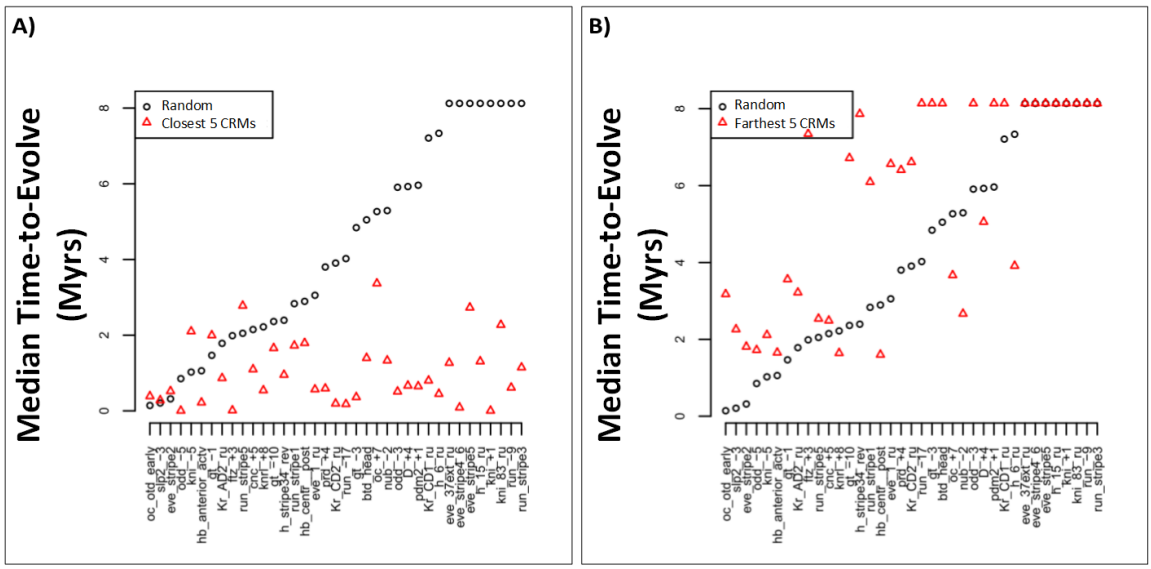


Figure 7.4: Effect of initial sequence on time-to-evolve estimates. (A) Simulations from initial sequence that drives a pattern similar to the target pattern (see Methods). Shown in red triangles are median time-to-evolve estimates (y-axis) for each target pattern (x-axis) under this modified simulation scheme; results from standard simulations where initial sequences are random are shown in black circles for comparison. A Paired Student's T-Test comparing the two sets of median time-to-evolve estimates yields a p-value of 2.78×10^{-8}. **(B)** Similar to (B) except that the red triangles represent simulations from an initial sequence that drives a pattern dissimilar to the target pattern (see Methods). This results in significantly greater time-to-evolve estimates, with p-value: 7.65×10^{-4} (Paired T-test).

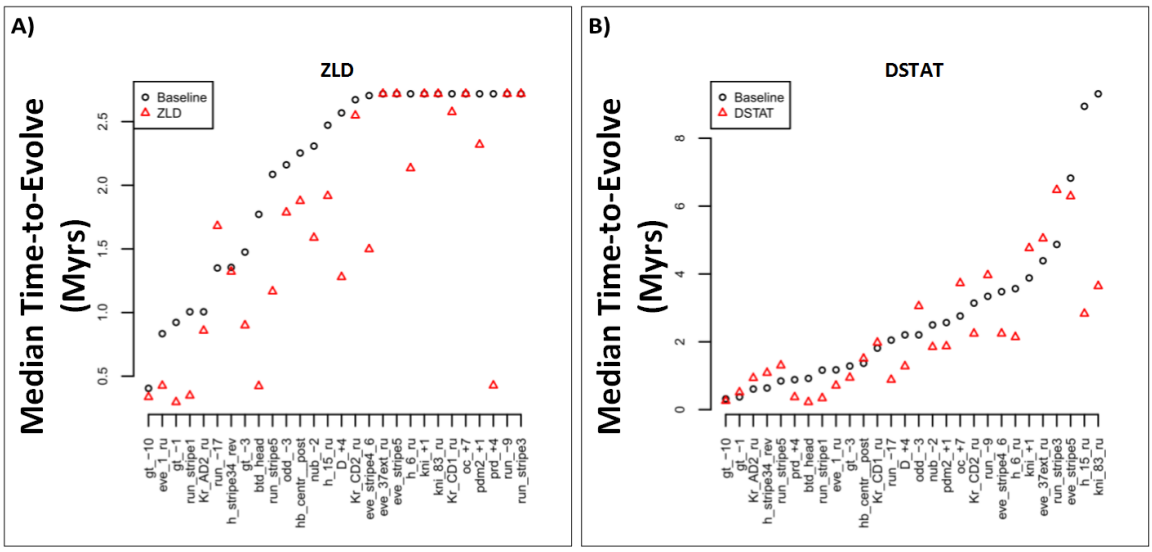


Figure 7.5: The effect of uniformly expressed activators on time-to-evolve for each target pattern.

(A) Comparison of time-to-evolve estimates between the baseline model and a model that includes ZLD as a uniform activator. Expression patterns are sorted based on time-to-evolve estimates from the baseline model. **(B)** Comparison of time-to-evolve estimates between the baseline model and a model that includes DSTAT as a uniform activator.

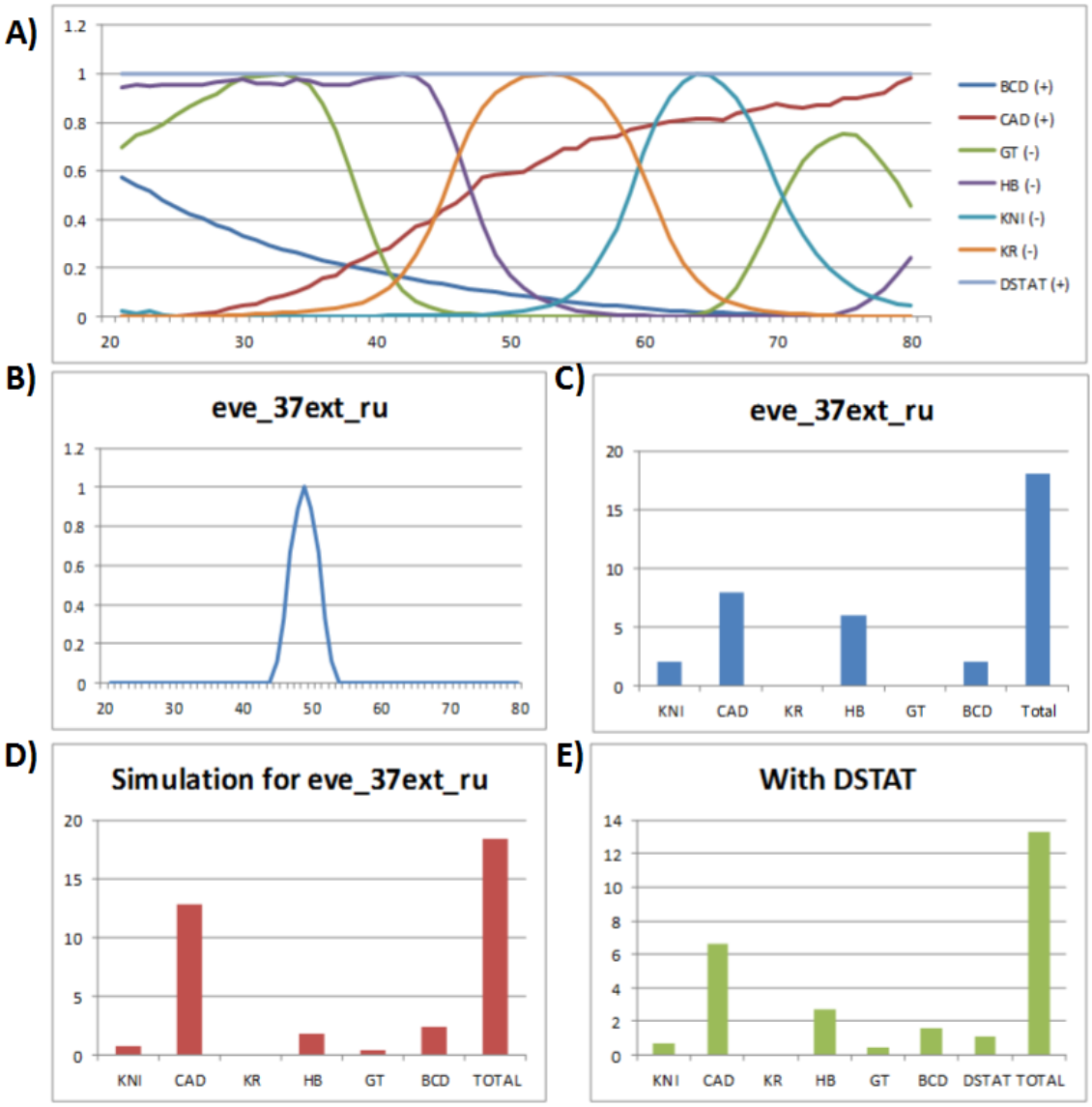


Figure 7.6: How a uniformly expressed activator affects combinatorial gene regulation. (A) Concentration profiles of seven TFs across the AP axis. Activators are indicated with a (+) and repressors with a (-). **(B)** Target expression pattern for the CRM 'eve_37ext_ru'. **(C)** Number of sites present in the eve_37ext_ru CRM in *D. melanogaster*, for each of six TFs (other than DSTAT). Sites are called at relative strength of 0.25 following the procedure described in Methods. **(D)** Number of sites for each TF in evolved CRMs, averaged over all simulations using the baseline model. **(E)** Same as **(D)**, for simulations using the alternative model that includes DSTAT.

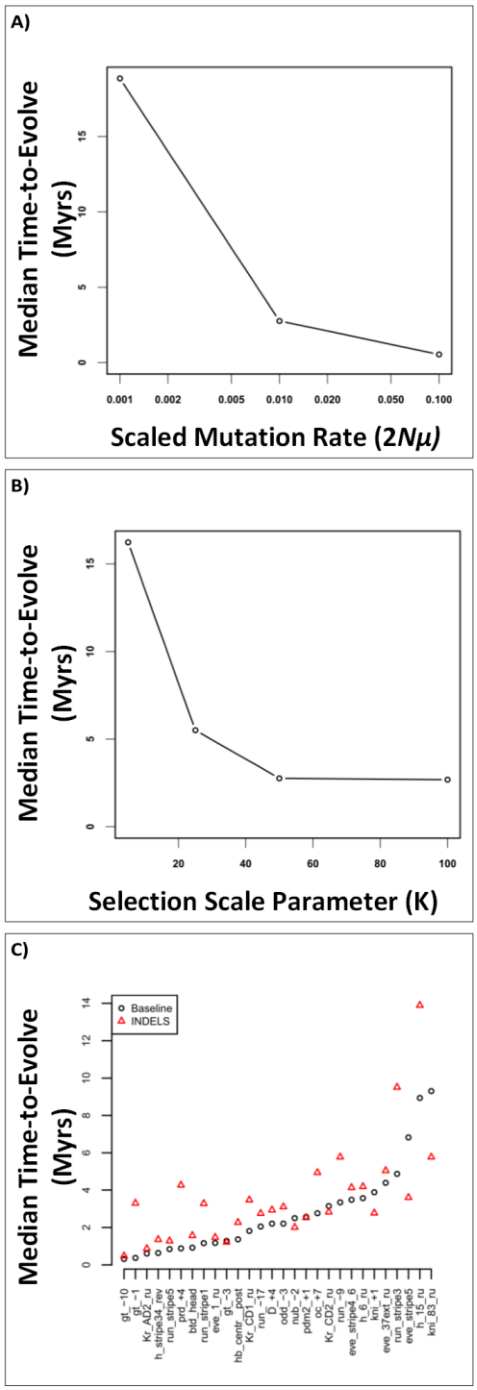


Figure 7.7: Sensitivity of time-to-evolve estimates to simulation parameters. (A) Sensitivity to the scaled mutation rate ($2N\mu$). Shown are the average time-to-evolve (median of all simulations for a pattern, averaged over 28 target expression patterns) for three values of $2N\mu$. **(B)** Sensitivity to selection-scale parameter (K). **(C)** The effect of indels. The plot shows time-to-evolve estimates (y-axis) for each of the 28 target expression patterns (x-axis) for an evolutionary model without insertions or deletions (black circles) and an evolutionary model that includes indels (red triangles). Adding indels significantly increases time-to-evolve estimates (P-value = 0.0002).

Table 7.1: Average pairwise distance between CRMs evolved *in silico* for the same expression pattern (d_{intra}) and for distinct patterns (d_{inter}). Patterns marked with a † superscript are those where the real CRM falls outside of the cluster of simulated CRMs.

Target pattern	d_{WT}	d_{intra}	d_{inter}	d_{inter}/d_{intra}	$\mathbf{d_{inter}/d_{WT}}$
h_15_ru†	6.13	2.36	4.66	1.97	**0.76**
kni_83_ru†	4.35	1.67	3.76	2.25	**0.87**
h_6_ru†	3.57	0.86	3.71	4.34	**1.04**
Kr_CD2_ru†	3.44	1.26	3.61	2.86	**1.05**
run_stripe3†	4.21	2.20	4.52	2.05	**1.07**
eve_37ext_ru†	4.19	1.68	5.03	3.00	**1.20**
kni_+1	2.62	1.36	3.54	2.61	**1.35**
D_+4	2.54	1.10	3.43	3.11	**1.35**
Kr_CD1_ru	2.40	0.79	3.37	4.29	**1.41**
run_-9	2.92	1.48	4.13	2.79	**1.41**
gt_-3	2.39	1.08	3.38	3.13	**1.41**
gt_-1	2.26	0.73	3.25	4.48	**1.44**
odd_-3	2.33	1.12	3.59	3.19	**1.54**
pdm2_+1	2.10	1.10	3.30	3.00	**1.57**
eve_stripe4_6	2.23	1.03	3.63	3.53	**1.63**
nub_-2	2.02	1.11	3.40	3.07	**1.68**
oc_+7	1.92	1.15	3.44	3.00	**1.79**
h_stripe34_rev	2.44	1.31	4.55	3.48	**1.87**
run_-17	1.81	1.15	3.86	3.36	**2.13**
gt_-10	1.51	0.73	3.68	5.03	**2.43**
hb_centr__post	1.48	1.05	3.59	3.41	**2.44**
btd_head	1.51	0.65	4.01	6.15	**2.67**
run_stripe5	1.19	1.14	3.40	2.97	**2.86**
prd_+4	1.16	0.56	3.51	6.23	**3.01**
Kr_AD2_ru	1.03	1.08	3.16	2.92	**3.07**
eve_stripe5	1.07	1.07	3.59	3.35	**3.35**
run_stripe1	1.04	0.61	4.24	6.90	**4.06**
eve_1_ru	0.81	0.40	3.77	9.37	**4.65**

7.3 Methods

7.3.1 Procedure for selecting expression patterns for simulation

To select the 28 patterns used in our experiments reported in Figures 1, 2, 3, 5 and 7, we repeated the following procedure on each of 37 expression patterns predicted by GEMSTAT (denoted by 'EP'),noting that these are the same expression patterns whose evolution was previously modeled using PEBCRES (20). First, we generated a set of random sequences that are fed into our fitness function. The fitness of each sequence was calculated using EP as the target expression pattern. If the average fitness of the random sequences was above 0.2, EP was not considered further. In other words, we chose to work with 28 expression patterns for which a random sequence has low fitness. At the same time, we know that the selected expression patterns are "achievable" in our framework since in each case we have a real CRM for which GEMSTAT predicts that expression pattern. We refer to each of these expression patterns by the name of the CRM from *D. melanogaster* that generated the pattern. The 28 expression patterns used for our experiments are shown Supplementary Figure A.1 on Appendix A.

7.3.2 Procedure for selecting starting sequences

By default, our simulations begin with a random sequence. In section "Dependence on initial conditions, and the possibility of exaptation", we report on two sets of experiments where the initial sequence was not random. Here, we first calculate the correlation between the expression patterns of every pair of CRMs and simulate the evolution of sequences targeting the expression pattern of one CRM while initializing the population with the sequence of another CRM. In one set of experiments we evolved an expression pattern from a sequence that already drove a similar pattern, by choosing the initial sequence for each simulation randomly from one of the 5 CRMs whose expression is most correlated with the target

expression pattern. In the other set of experiments we evolved a pattern from a sequence that drove a very different pattern, by choosing the initial sequence randomly from one of the 5 CRMs whose expression is most anti-correlated with the target expression pattern.

7.3.3 Procedure for comparing different GEMSTAT model specifications

Our evolutionary simulations require a model of CRM function, which is provided by GEMSTAT, to help define the fitness function. The model in GEMSTAT can be specified to include or exclude a specific regulator, and once a regulator is added to the model, all parameters are learnt from appropriate training data. Our goal was to compare evolutionary simulations made with two different specifications of the GEMSTAT model: one that includes a ubiquitous activator and one that does not. However, there are a couple of concerns to be addressed before such comparisons can be made.

Recall that the baseline model, i.e., the model without the universal activator, is used to define the target expression pattern of a simulation. Specifically, as noted in *Uniformly expressed activators can speed up emergence of CRMs*, we take a *D. melanogaster* CRM, use the baseline model to predict its expression pattern (say '*T*'), and use this pattern as the target of a PEBCRES simulation. This ensures that the simulation is using a fitness function such that there is at least one sequence with perfect fitness. Now, we may train a new GEMSTAT model (say 'M_U') that includes the universal activator, and perform simulations using this new model to define fitness. These simulations must target the same expression pattern (*T*) as before, to make claims about the role of the ubiquitous activator in shaping the evolutionary dynamics. However, there is no guarantee that there exists a sequence with perfect fitness when using the new model M_U. That is, there may not exist a sequence for which model M_U predicts expression pattern *T* exactly. This makes the comparison unfair, since the existence of a perfectly solution is only guaranteed for one of the models. An alternative is to run the simulations both simulations (with baseline model or with M_U) with a new target expression pattern (say *T'*), set to be the prediction of M_U on the *D. melanogaster* CRM. This guarantees that the simulations with M_U as fitness function can in principle find a sequence with perfect fitness, but the new pattern *T'* may require the use of the ubiquitous activator and simulations

with the baseline model may not have any chance of finding the perfectly fit CRM. Our hypothesis is that ubiquitous activators reduce the time necessary to evolve certain expression patterns, even if the same pattern might have been evolved without utilizing the ubiquitous activator. To test this hypothesis we need a setup where the fitness function includes regulatory input by a ubiquitous activator but the latter is not necessary for a solution to have high fitness. To this end, we use the experimental setup described below:

1) Start with the baseline GEMSTAT specification (TFs: BCD, CAG, GT, HB, KNI, KR; self-cooperativity for BCD, CAD, KNI).
2) Train on all 37 CRMS an alternative GEMSTAT specification that includes a ubiquitous activator (either DSTAT or ZLD). All other assumptions of the baseline model are maintained. The alternative specification should be trained to match the expression patterns predicted by the baseline model. This will result in an alternative model (say M_U) whose predicted expression for each of the 37 CRMs is very close to the predictions from the baseline model.
3) For each CRM, merge the predicted expression patterns from the baseline model and from M_U by taking their average. The merged expression pattern is thus equally "achievable" by either model.
4) Repeat the experiment to determine median time-to-evolve per CRM using the baseline model as the fitness function, but targeting the merged expression pattern. This is only done for the 28 CRMs shown in Figure 1.
5) Repeat the experiment to determine median time-to-evolve per CRM using M_U the fitness function, again targeting the merged expression pattern.
6) Compare median time-to-evolve per CRM for simulations from steps 4 and 5.

Our experimental set up still does not guarantee that there exists a solution with fitness of 1 during simulations, but manual inspection assured us that in each simulation, whether it uses the baseline model or M_U, there is at least one sequence with fitness ~1 with respect to the target expression pattern defined as above. We repeated the above procedure for two alternative models M_U, the first one including ZLD and the second including DSTAT as the additional ubiquitous activator.

7.4 Summary and conclusion

We used the evolutionary simulation framework of PEBCRES, from our previous work [CITE], to study what it might take for a functional CRM to evolve, first asking how long this may take under strong selection, and then investigating various factors that may influence the estimated time-to-evolve. These questions have been addressed in various ways by other authors before us, ,for instance by Stone and Wray (2001), MacArthur and Brookfield (2004) and Durrett and Schmidt (2007, 2008). Early approaches to this question focused on the independent evolution of single binding sites (Stone and Wray 2001; Durrett and Schmidt 2007), pairs of binding sites (Durrett and Schmidt 2008) or simple CRMs composed of a single TFs (MacArthur and Brookfield 2004). However, questions regarding CRM evolution assume additional complexity due to the diverse mechanisms and combinatorial nature of gene regulation, which have not been adequately addressed in previous work. In recent work, we developed the PEBCRES evolutionary framework to bridge this gap, and used it to accurately model the evolutionary dynamics of binding sites within CRMs under strong negative selection for a fixed regulatory function (Duque et al. 2013). The success of that work encouraged us to explore here a complementary aspect of CRM evolution – that of emergence of a new CRM under strong adaptive forces.

We estimated that CRMs that exhibit the combinatorial complexity associated with early developmental enhancers (specifically, those involved in anterior-posterior patterning in *Drosophila* embryos) can emerge on fairly short time scales, of the order of few millions of years, even when starting from random sequences of little or no functional ability. A recent study [CITE doi:10.1038/ng.3009] used massively parallel enhancer screens (STARR SEQ [CITE]) to find that hundreds of novel CRMs have emerged on the scale of ~10 Myrs, lending credibility to our theoretical findings. While we are not aware of other previous studies reporting time-to-evolve estimates for CRMs, it is worth noting that Durrett and Schmidt (2007) estimated that an 8 bp long ‘fuzzy’ binding site (one mismatch allowed) might emerge in the human population on a timescale of 60,000 years. A CRM evolved in our simulations for the ‘gt_-10’ pattern, for example, has about 7 binding sites on average, and takes about 0.3 million years to evolve. This

agrees roughly with an extrapolation from Durrett and Schmidt (2007) whereby the time for 7 binding sites to emerge should about 0.42 million years, assuming sites are not lost and sites emerge sequentially. This is a ballpark comparison, since the two estimates are for human and fruitfly populations respectively and contingent upon different assumptions about a binding site's information content.

Here, CRMs were evolved in silico to drive pre-determined expression patterns along the A/P axis. CRMs arising from different simulations for the same target expression pattern tended to cluster strongly in terms of site composition, with distinct expression patterns defining distinct clusters of "fit" enhancers. Importantly, we noted evolved sequences to be similar in site composition to the real *D. melanogaster* CRMs associated with their respective patterns, thus demonstrating agreement between model-based evolutionary simulations and real data. The few exceptions from this general trend were also illuminating, with the evolved CRMs being significantly more parsimonious than their real counterparts, leading to speculations about a more complex evolutionary history of those real CRMs or about missing regulatory mechanisms in the PEBCRES/GEMSTAT framework. This latter point deserves special mention as missing regulatory mechanisms can shade the findings of simulation-based studies, as was demonstrated in this thesis. For instance, we note that the A/P patterns used as targets in our simulations lack terminal aspects – we only considered the regulatory function of a CRM in the range 20% - 80% egg length. This may lead to underestimates of time-to-evolve CRM for certain patterns. Proper modeling of these CRMs requires that the underlying fitness function, specifically GEMSTAT, use additional TFs, some of which are not known (He et al. 2010). Additionally, previous work on GEMSTAT (He et al. 2010) and PEBCRES (Duque et al. 2013) have produced careful estimates for many of the free parameters used in this work by modeling expression patterns that excluded the terminal ends. These were two major reasons why we decided to exclude the terminal ends of the embryo from our analysis.

We noted up to a 30 fold variation in the time to evolve CRMs for different expression patterns, naturally raising the question: what causes this variable time-to-evolve? The flexibility inherent in the PEBCRES framework (as opposed to a purely analytical framework) allowed us to explore

different aspects of the evolutionary process and regulatory mechanisms, and how they might affect emergence time of CRMs. For example, we asked how these times might be affected if a CRM, instead of evolving from genomic background, arose from a sequence that already drives some expression pattern. Unsurprisingly we found that if the two expression patterns (that driven by the original sequence and the target pattern) are highly similar, the time to evolve a CRM is greatly reduced; however, perhaps more interestingly, the emergence time can also be significantly greater if the expression patterns are very dissimilar. Dependence of evolution times on initial sequences has been proposed in previous work. For instance, MacArthur and Brookfield (2004) argued that the time to evolve a CRM that drives a certain level of activation by a TF may be influenced by the CG-content of the initial sequence.

As another example of factors affecting time-to-evolve, we found that ubiquitous activators, which are not by themselves capable of patterning a target gene, may work with other TFs and reduce the time to evolve a CRM. We speculate that this may be due to two complementary reasons: 1) ubiquitous activators provide alternative solutions to the underlying combinatorial optimization problem of finding a "fit" CRM, and 2) ubiquitous activators reduce the number of binding sites necessary to create certain expression patterns, and thus the number of steps (mutations) needed to find a fit solution. Both situations are expected to reduce the time to find one such solution, as per theories of evolutionary computation (Goldberg 2002).

Our simulations suggest that CRMs with more combinatorial regulation (measured by the number of TFs with sites in the *D. melanogaster* CRM for the same pattern) should take longer to evolve. Perhaps more surprisingly, we noted that the binding site content for a particular TF – Hunchback (HB) – is one of the strongest predictors of time-to-evolve values. It is possible that this points to shortcomings of our simulation framework. We noted that the HB motif can be characterized as a poly-T repeat (Supplementary Figure A.3 on Apendix A). Such repeat patterns might be easily created through mutational mechanisms that we have not modeled adequately in PEBCRES. Moreover, there is evidence that HB might play dual roles of activator and repressor depending on the regulatory context. The absence of this mechanism in our GEMSTAT-based fitness function may be related to the strong correlation noted above, and

illustrate more generally how evolutionary modeling may lead us to closer examination of mechanisms encoded in cis-regulatory sequences.

We also found that anterior expression patterns were quicker to evolve sequences for than posterior patterns. However, this observation is likely a consequence of the already mentioned influence of HB site counts. Noting that HB is modeled as a repressor and is largely expressed in the anterior end of the embryo, anterior expression correlates with lesser site content for HB, which in turn correlates with shorter time-to-evolve values.

Our application of PEBCRES to understanding the evolution of CRMs can be extended in several ways. For example, our model could be used to shed light on shadow enhancers (Perry et al. 2010; Barolo 2012), by using the GEMSTAT-GL model of locus-level modeling for regulatory function prediction instead of the GEMSTAT model of enhancer function. Other avenues of future exploration include understanding the effect of indirect activators (Kanodia et al. 2012), the effect of local duplications (Sinha and Siggia 2005) on time-to-evolve estimates, exploring the robustness of evolved CRMs to fluctuations in input TF concentrations (Pujato et al. 2013) and how such robustness might evolve (Wagner 2005), and understanding how evolvability (Wagner and Altenberg 1996; Wagner 2005) affects the architecture of cis-regulatory sequences and how it evolves in the first place (Wagner and Altenberg 1996).

8 Concluding remarks

We have presented a new methodology to model the evolution of regulatory sequences. Our approach consists on using a state-of-the-art sequence-to-expression model to predict the effect of mutations to the regulatory sequence. We show that our model is capable of accurately modeling the evolution of 37 CRMs from the AP system among 12 Drosophila species.

We also used our model to show that the widespread phenomenon of homotypic clustering is, at least to a certain extent, expected as a consequence of the evolutionary process, even when no fitness advantage exists. Next, we use our model to gather evidence that CAD molecules interact homotypicaly, a hypothesis that was later tested and confirmed using experimental techniques. Finally, we discuss how our approach can be used to learn more about the evolutionary process itself and about how evolvable an expression pattern is. Our methodology has allowed us to gain new insights into two important biological processes: the regulation of gene expression and the evolution of the sequences involved in this regulation. The results achieved this far demonstrate relevance, both biological and computational, of our model.

Appendix A – Supplementary material for chapter 7

Supplementary Note A.1

To illustrate an issue with evolutionary models that operate at the level of individual binding sites, we can look at compensatory mutations. This is the phenomenon where a deleterious mutation is followed by a mutation elsewhere, typically in the same CRM, that compensates for the deleterious effect of the first mutation. Initially, the prevalent theory was that the deleterious mutation would fix due to genetic drift and, subsequently, the compensatory mutation would fix due to positive selection. This theory predicts such turnover events to be more common on smaller populations (such as vertebrates compared to invertebrates), since drift is accentuated is smaller populations. However, sequence comparison data shows that turnover events are in fact more common in larger invertebrate populations (1), where the effect of genetic drift is expected to be more limited.

Such a paradox could not be explained by modeling the evolution of binding sites as independent events, and was notably addressed by (1). By explicitly modeling the evolution of pairs of binding sites, instead of single binding sites, they identified the cause of this paradox to be the phenomenon called *stochastic tunneling* (2), in which the second (compensatory) mutation happens before the first (deleterious) mutation is fixed. The result of including *stochastic tunneling* in their population genetics model is that turnover events become more common for populations of size similar to invertebrate populations (1).

Supplementary Note A.2:

Our goal was to examine the adaptive evolution of sequences that can drive the anterior-posterior patterns experimentally observed for real *D. melanogaster* CRMs. However, for reasons explained in the main text, we designated as our target expression patterns the patterns predicted by GEMSTAT for these real CRMs, rather than the actual expression patterns driven by them. Therefore, it was worth asking if the predicted expression patterns used as targets of evolution are good approximations of the real patterns. Supplementary Figure X1

shows that many of the predicted expression patterns closely match their real counterparts and that for many others the important characteristics of the patterns are well captured; however, visual examination indicates that for some patterns the agreement is much poorer. Similarly, one may use an objective measure to score the agreement between the predicted and real patterns. The problem with either approach is that it is hard to determine an appropriate threshold for the minimum agreement between real and predicted patterns. For this reason, we decided not to filter any pattern on account of its quality of fit to the real pattern, but rather consider all of the 37 patterns modeled in (3). Additionally, we note that in most of our results, the patterns with worst agreement between predicted and real patterns (e.g., hb_centr_post or Kr_AD_ru) are mostly median and therefore do not significantly influence any of our conclusions.

Supplementary Note A.3

We limited the number of generations simulated to 100,000. However, since our simulation time is limited there is a chance that certain simulations might never yield an individual with fitness above the threshold (For example, see Supplementary Figure Xy). If this is the case we set the time-to-evolve for that simulation to be the maximum time allowed for the simulation (100000 generations in this particular experiment, or equivalently, about 8 Million Years). Since we examine the median time-to-evolve from many simulations for each target pattern, this has no effect as long as more than 50% of the simulations for a CRM yield an individual with fitness above threshold. If less than 50% of the simulations yield a fit individual, the median time for the CRM will be underestimated as the maximum time allowed in our simulations.

Supplementary Note A.4

Our goal was to summarize a CRM in a way that allows us to compare two CRMs for similar cis-regulatory logic, even if they are not evolutionarily related, i.e., did not evolve from the same ancestor sequence. One reasonable way to summarize a CRM, commonly adopted in the literature, is to count binding sites (above a threshold) for each relevant TF. However, this method ignores information regarding, for example, the position of the binding sites. This

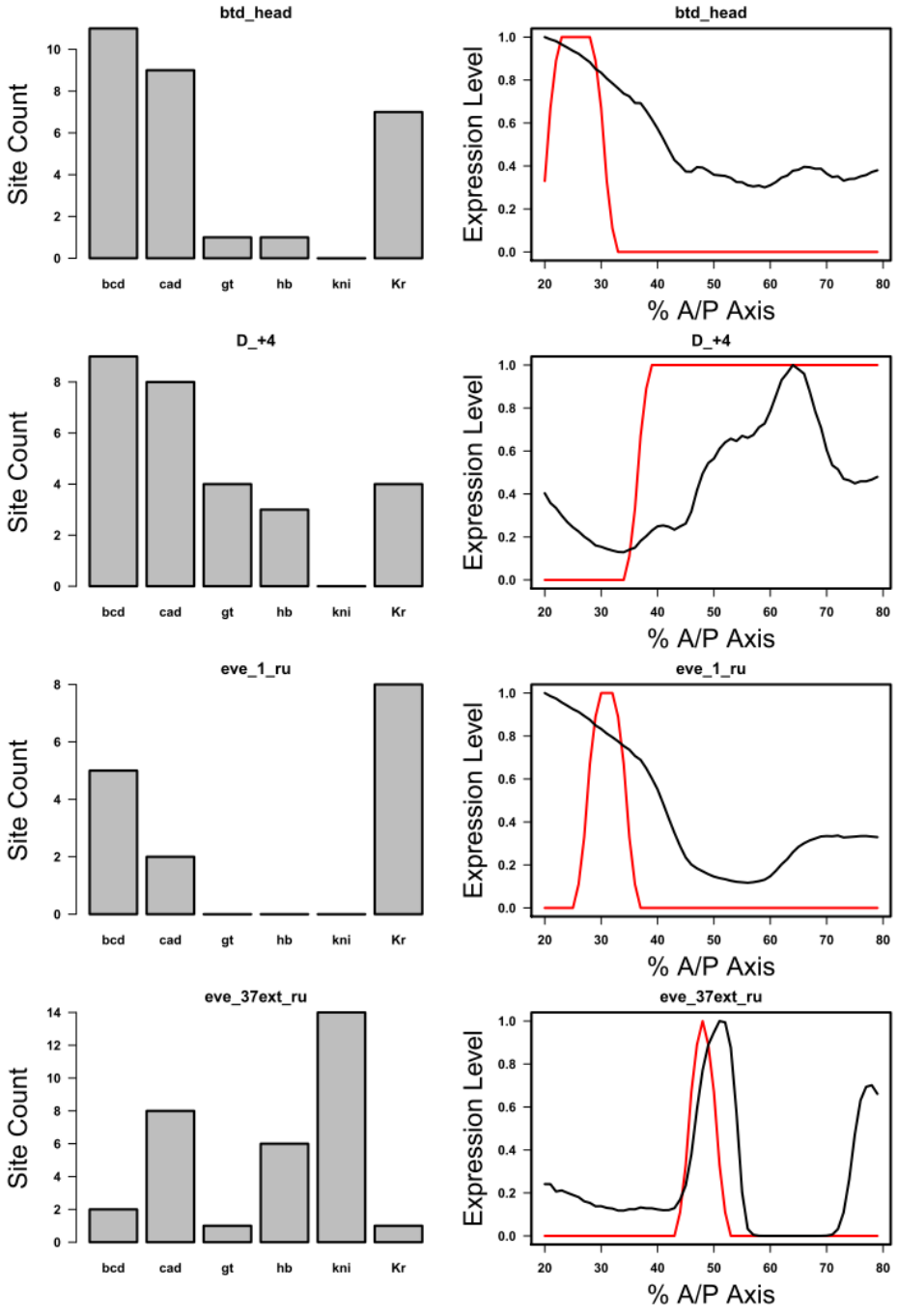
information is especially important for cooperative pairs of binding sites and short range repressors. Additionally, as shown in our previous work on adaptive evolution of BCD-driven CRMs (4), evolution might produce CRMs with a wide range of binding site counts even for simple expression patterns. (It is possible to create similar expression levels using several weak binding sites or fewer stronger binding sites.)

The estimated TF occupancy provides a better alternative to the site count above a threshold. We have demonstrated in our previous work (4) that the occupancy of a TF, a "weighted" binding site count that takes into consideration the strength of each binding site, has a relatively smaller spread of values in evolved CRMs for a simple target pattern. Additionally, the occupancy measure defined in (4) is based on the statistical thermodynamics (5) formulation of GEMSTAT (3), and accounts for some position specific phenomena such as cooperative interactions. Finally, the occupancy value is an integrated measure of strong and weak sites and does not require an ad hoc thresholds to define binding sites.

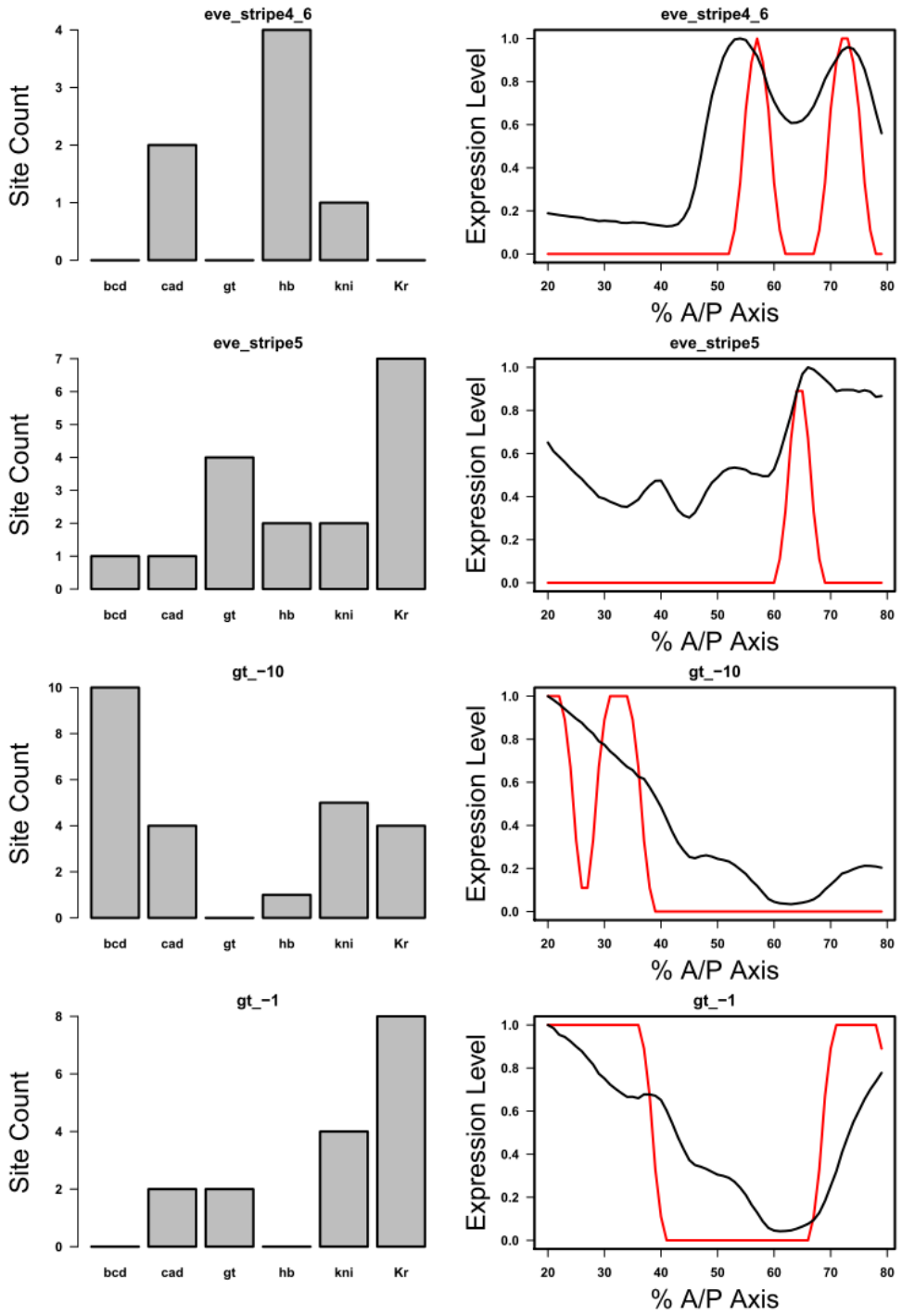
Supplementary Note A.5

The target pattern 'h_15_ru', whose simulation results least resemble the real *D. melanogaster* CRM (Figure 2F**), comprises two expression domains, one anterior and one posterior (See Supplementary Figure X1). To create such patterns the real *D. melanogaster* CRM uses binding sites for the activators BCD and CAD, to induce expression in the anterior and posterior domains respectively, as well as sites for the repressors KR, HB and GT (Supplementary Figure X1). Simulated CRMs for the same target pattern employ roughly the same strategy as the real CRM, but they utilize, on average, about half as many binding sites (estimated occupancy) as the real CRM, for KR, BCD and HB, and fewer GT sites also. The real h_15_ru CRM seems to harbor unusually many sites of these TFs, especially KR, when compared to other real CRMs. A similar phenomenon can be observed for 'kni_83_ru' (Figure 2E**), another target pattern with relatively large d_{WT} (Table 1**). According to GEMSTAT, the posterior domain expression driven by the real CRM is achieved through binding sites for the activator CAD and repressors HB, GT

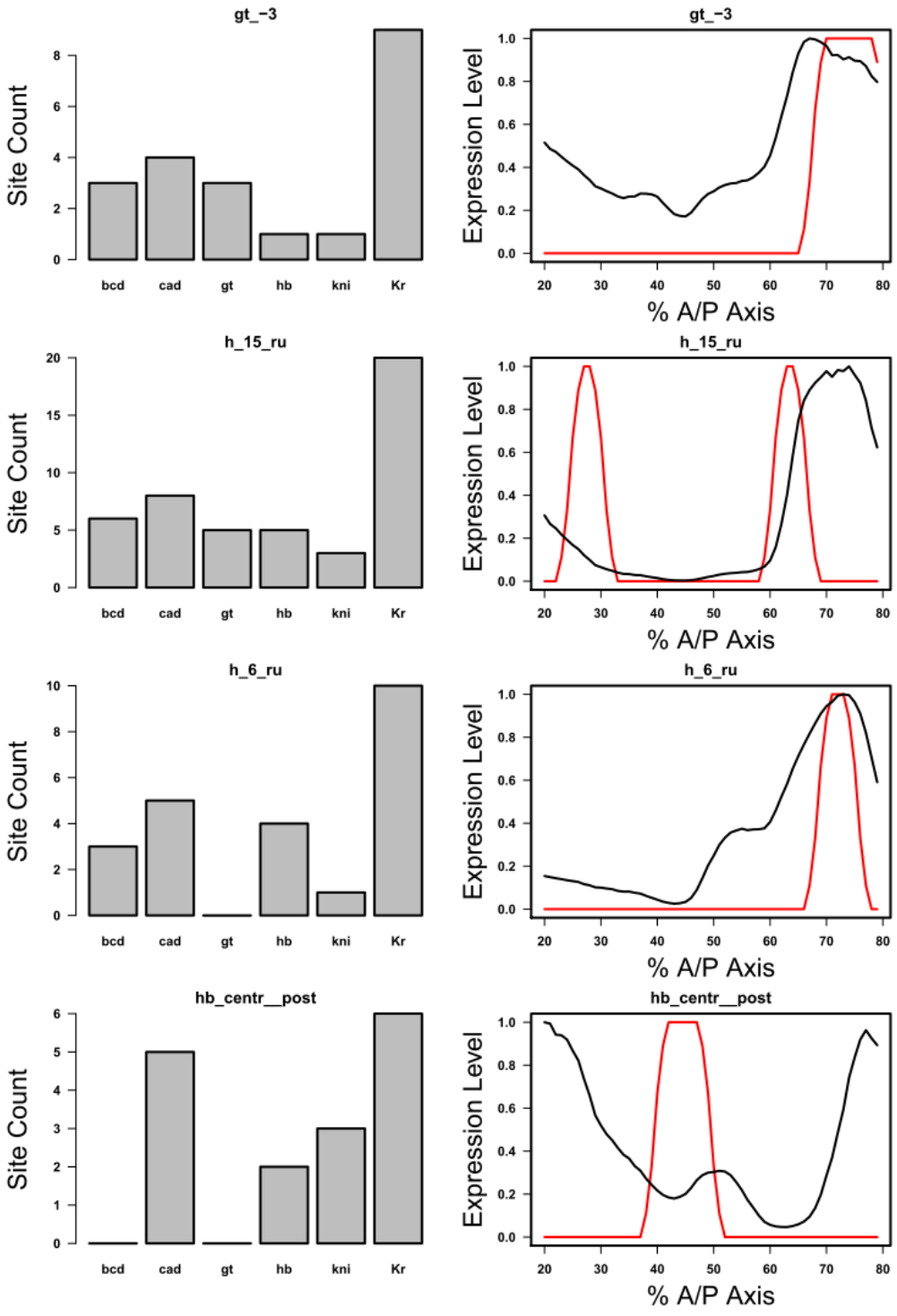
and KR. The simulated CRMs use the same combination of TFs but with fewer sites (lower estimated occupancy) for all involved factors.



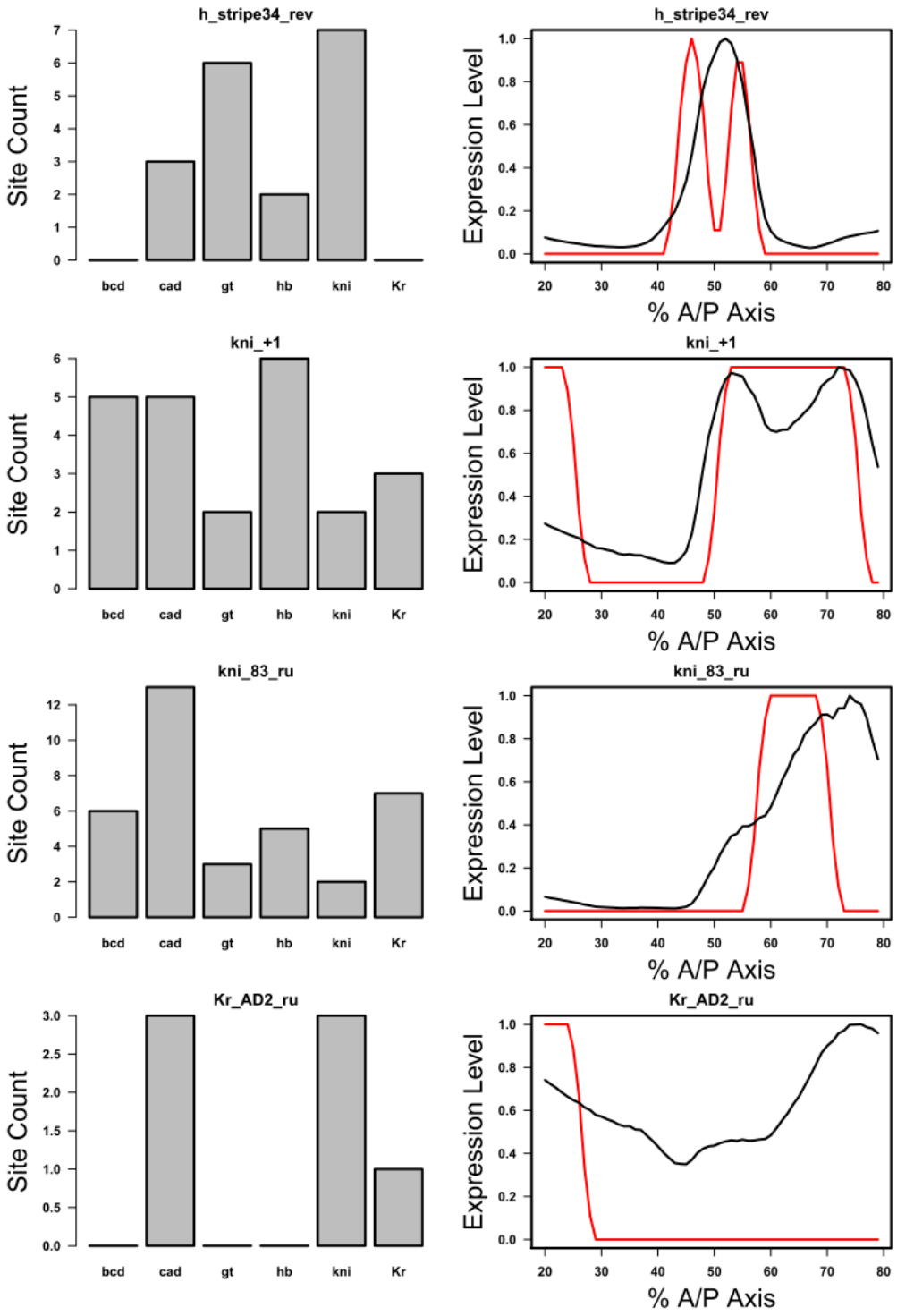
Supplementary Figure A.1 part 1 of 7



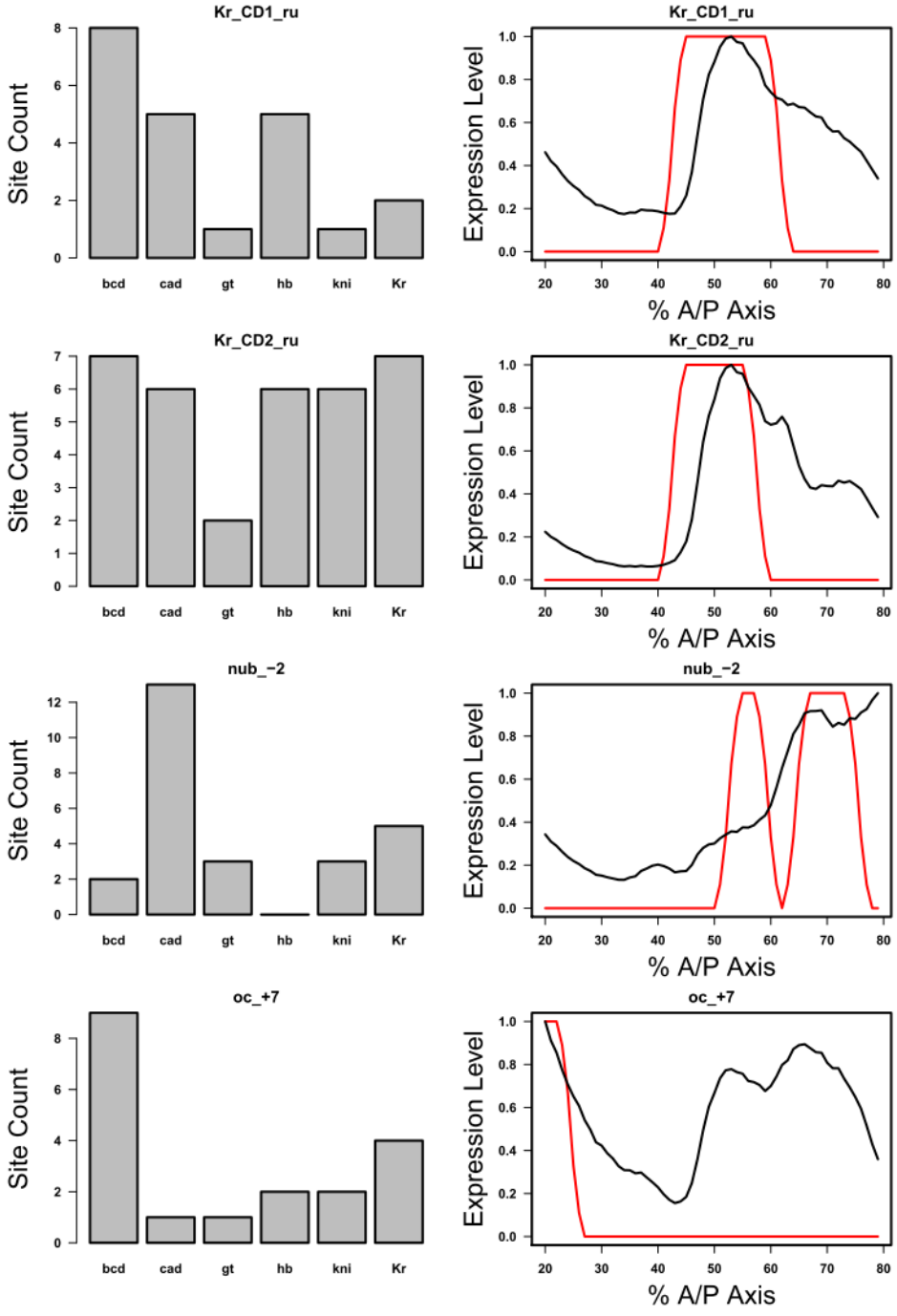
Supplementary Figure A.1 part 2 of 7



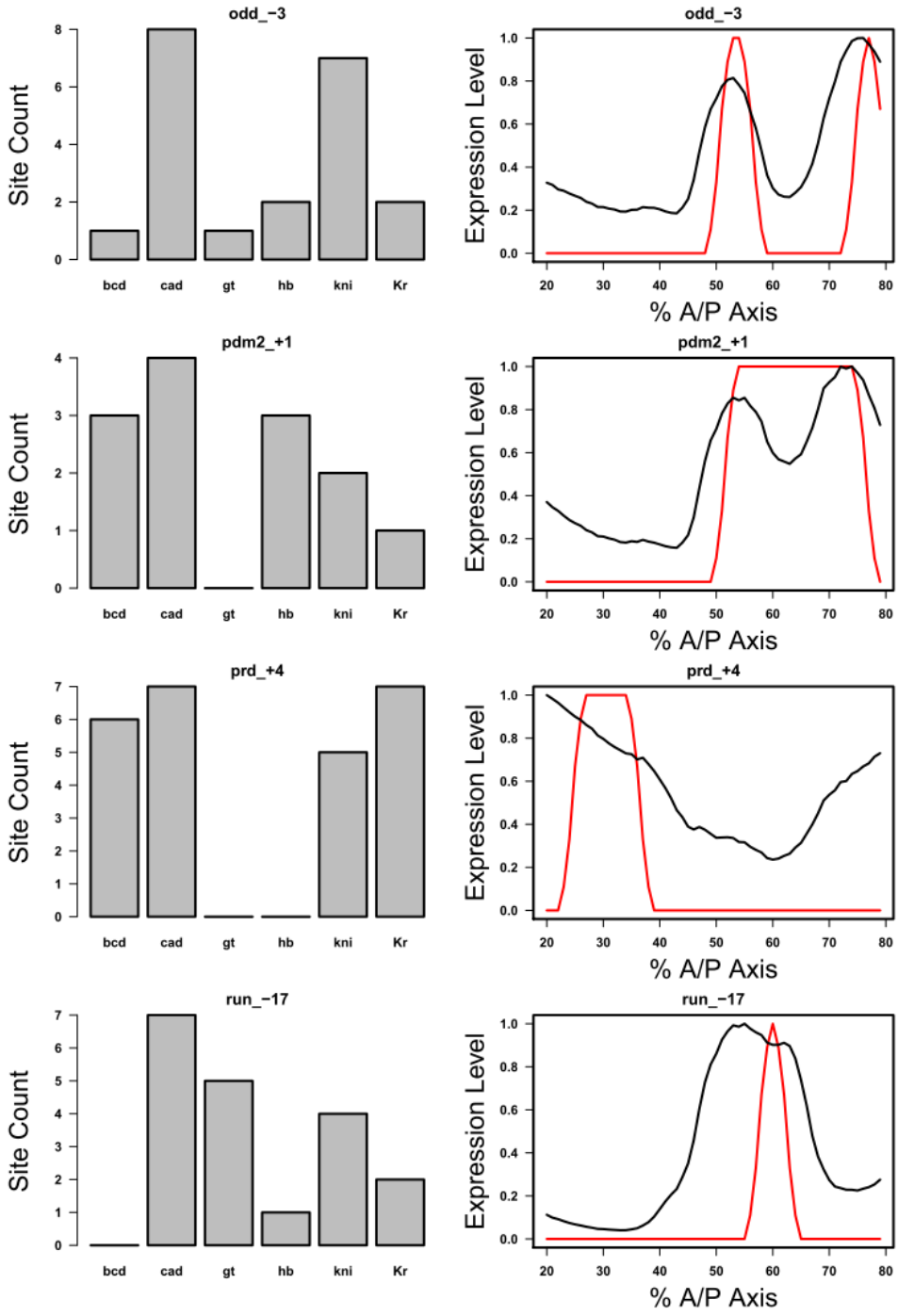
Supplementary Figure A.1 part 3 of 7



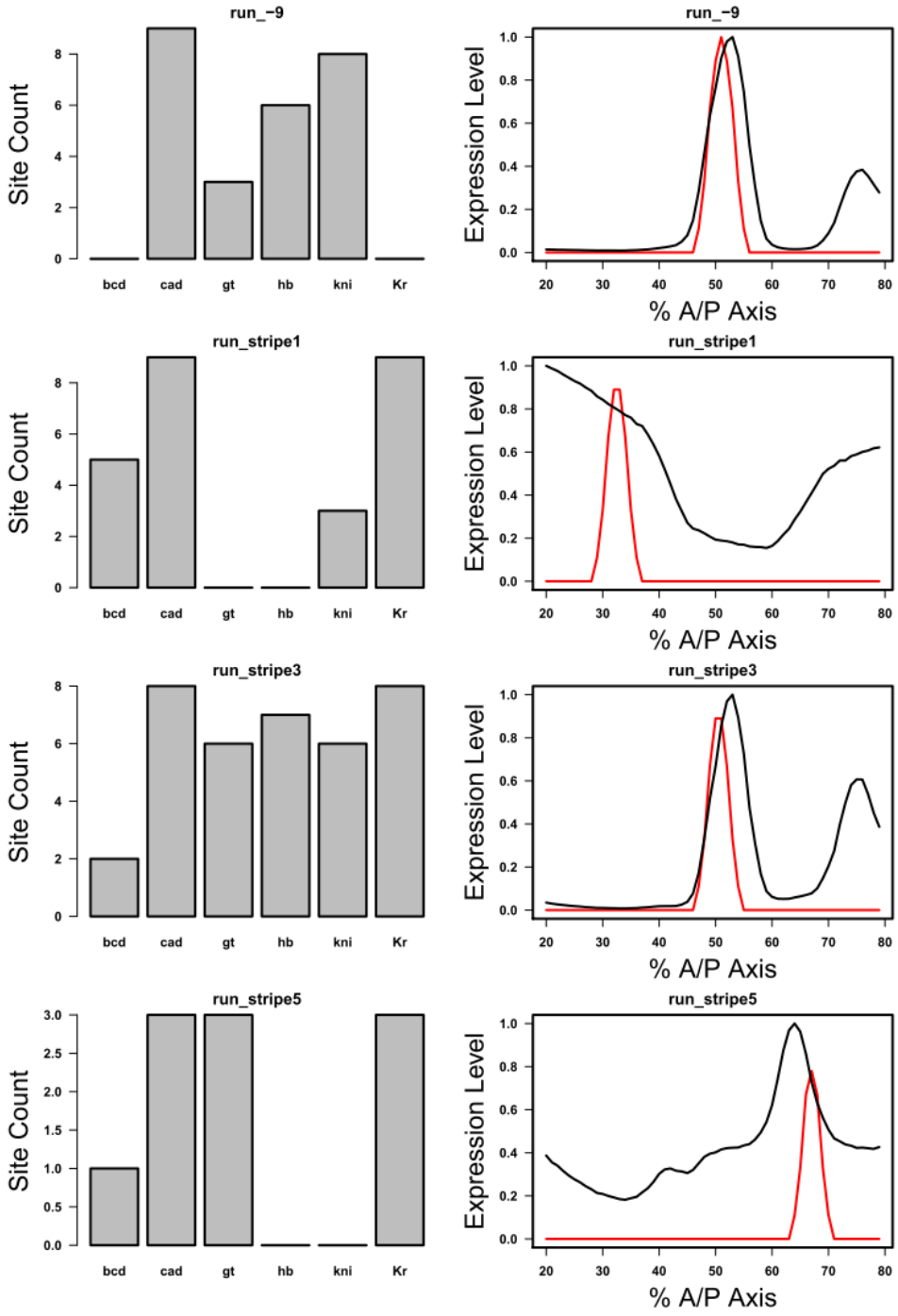
Supplementary Figure A.1 part 4 of 7



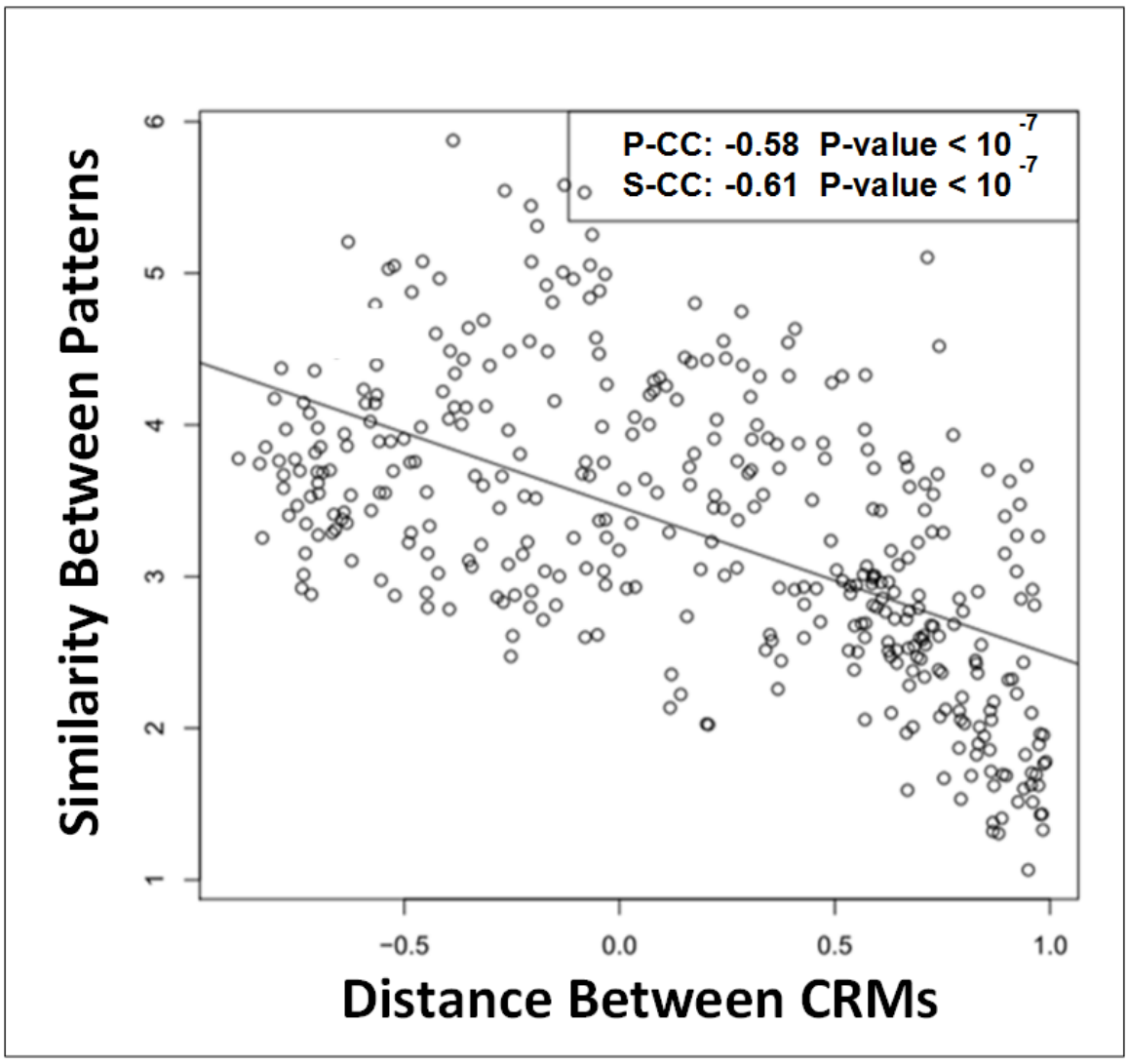
Supplementary Figure A.1 part 5 of 7



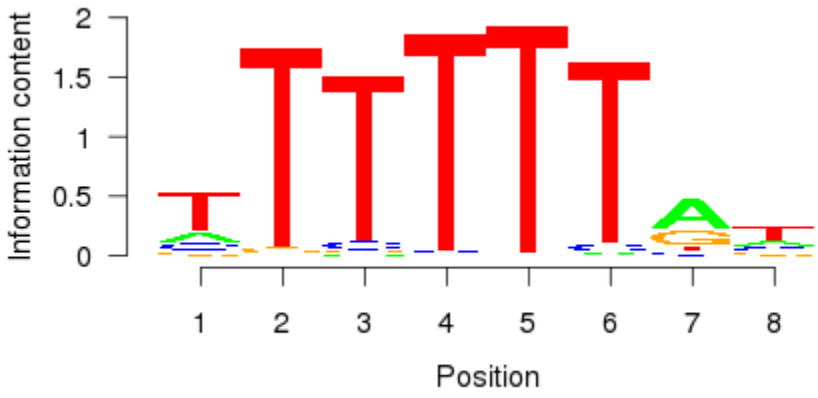
Supplementary Figure A.1 part 6 of 7



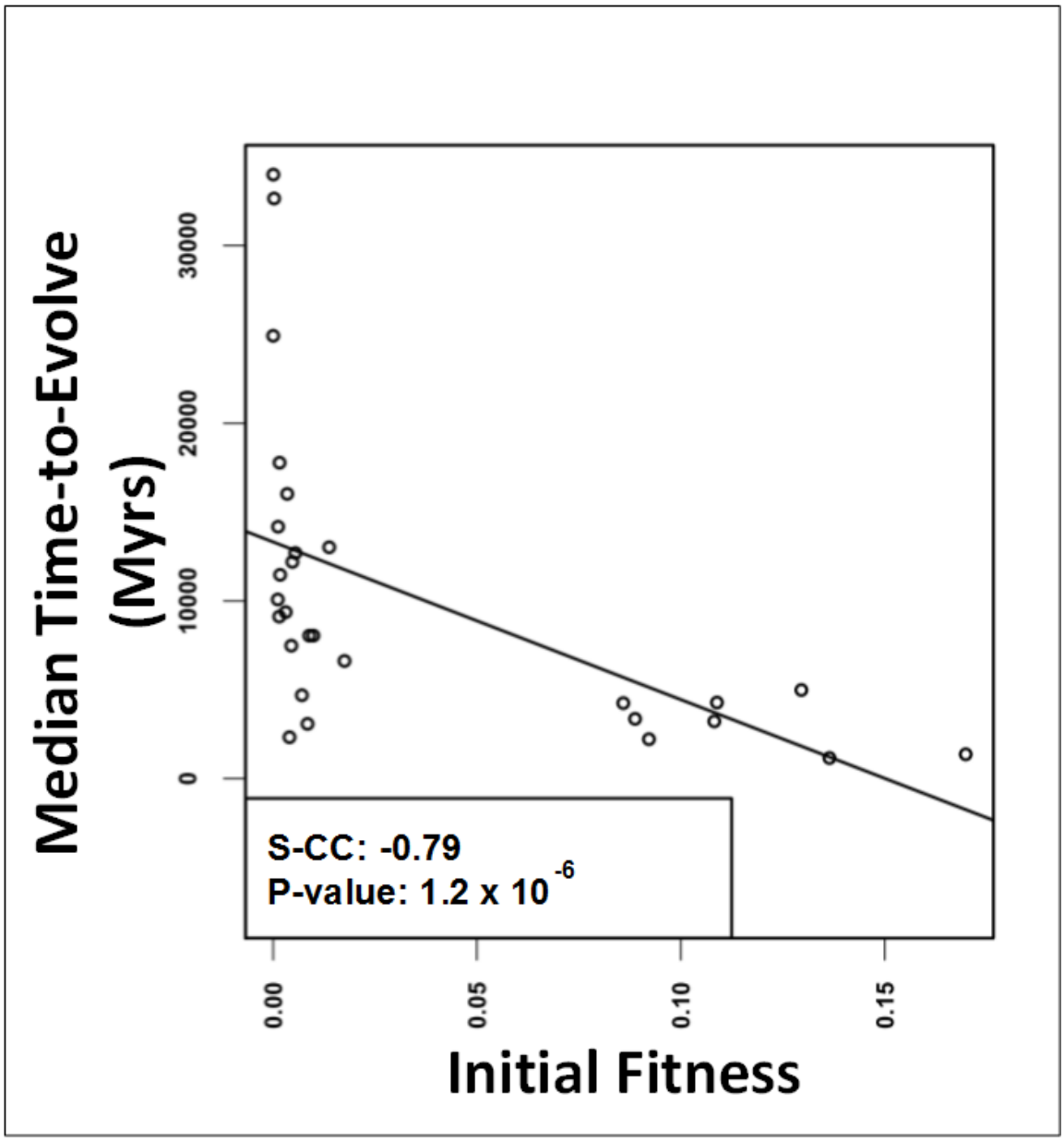
Supplementary Figure A.1: Predicted and real expression patterns. Each CRM is represented by two panels. The top panel shows the real expression pattern for each obtained from (3) (red line) and the expression pattern predicted for that CRM using GEMSTAT (black line). The bottom panel shows the number of binding sites for each of 6 TFs: BCD, CAD, GT, HB, KNI and KR.



Supplementary Figure A.2: Similarly expressed CRMs have similar occupancy vectors. A scatter plot relating the distance between a pair of CRMs in the six-dimensional space of TF occupancy values ('Distance between CRMs', y-axis) and the correlation coefficient between the expression patterns driven by those CRMs ('Similarity between patterns', x-axis). The correlation between the two variables is significant: Pearson Correlation Coefficient ('P-CC') = -0.58, P-value < 10^{-7}; Spearman Correlation Coefficient ('S-CC') = -0.61 P-value = < 10^{-7}. The best fit line is shown as a solid line.



Supplementary Figure A.3: The HB motif.



Supplementary Figure A.4: Scatter plot relating the average initial fitness across multiple simulations of a target pattern (x-axis) and the median estimated time-to-evolve for that pattern (y-axis). Correlation between the two variables is significant (Spearman's CC = -0.79, P-Value = 1.2×10^{-6}); however, this correlation can be attributed to the binding site content for HB in the *D. melanogaster* CRM associated with the target pattern (partial correlation is insignificant at P-value = 0.12, also see Figure 7.3(A)).

Supplementary Table A.1

Two-Way ANOVA table to access the significance of the reduction in time-to-evolve for all the CRMS when adding the TF DSTAT. The time-to-evolve is evaluated at fitness threshold of 0.64.

	Degrees of Freedom	Sum Squared Errors	Mean Squared Errors	F Value	P-Value
CRM	27	1.7595×10^{11}	6.5167×10^{9}	39.8220	$< 2.2 \times 10^{-16}$
Model	1	6.4579×10^{9}	6.4578×10^{9}	39.4622	4.1×10^{-10}
CRM x Model	27	1.7244×10^{10}	6.3867×10^{8}	3.9028	7.4×10^{-11}
Residuals	1934	3.1649×10^{11}	1.6364×10^{8}		

Supplementary Table A.2

Two-Way ANOVA table to access the significance of the reduction in time-to-evolve for all the CRMS when adding the TF ZLD. The time-to-evolve is evaluated at fitness threshold of 0.64.

	Degrees of Freedom	Sum Squared Errors	Mean Squared Errors	F Value	P-Value
CRM	27	1.8105×10^{11}	6.7054×10^{9}	36.5529	$< 2.2 \times 10^{-16}$
Model	1	8.6167×10^{9}	8.6167×10^{8}	4.6972	0.0303
CRM x Model	27	2.3111×10^{10}	8.5596×10^{8}	4.6660	3.5×10^{-14}
Residuals	1934	3.5478×10^{11}	1.8344×10^{8}		

Summary of abbreviations

AP – Anterior/Posterior patterning

BCD – Bicoid (transcription factor)

bp – Base pair

BTM – Basal Transcription Machinery

CAD – Caudal (transcription factor)

CRM – Cis-Regulatory Module (a.k.a. enhancer)

HB – Hunchback (transcription factor)

HMM – Hidden Markov Model

KNI – Knirps (transcription factor)

KR – Kruppel (transcription factor)

LLR – Log-Likelihood Ratio

LR – Likelihood Ratio

PEBCES – Predicted-Expression-Based CRM Evolution Simulator

PEBSES – Predicted-Expression-Based Site Evolution Simulator

PWM – Position-specific Weight Matrix

SS – Site Simulator (model from (Kim et al. 2009))

TF – Transcription Factor

TFBS – Transcription Factor Binding Site

TSS – Transcription Start Site

References

Akam M. 1987. The molecular basis for metameric pattern in the Drosophila embryo. Development 101: 1-22.

Anderson GM, Freytag SO. 1991. Synergistic activation of a human promoter in vivo by transcription factor Sp1. Mol Cell Biol 11: 1935-1943.

Arbouzova NI, Zeidler MP. 2006. JAK/STAT signalling in Drosophila: insights into conserved regulatory and cellular functions. Development 133: 2605-2616. doi: 10.1242/dev.02411

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339: 1074-1077. doi: 10.1126/science.1232542

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem 94: 890-898. doi: 10.1002/jcb.20352

Bakkali M. 2011. Microevolution of cis-regulatory elements: an example from the pair-rule segmentation gene fushi tarazu in the Drosophila melanogaster subgroup. PLoS One 6: e27376. doi: 10.1371/journal.pone.0027376

Balhoff JP, Wray GA. 2005. Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. Proc Natl Acad Sci U S A 102: 8591-8596. doi: 10.1073/pnas.0409638102

Ballestar E, Esteller M. 2008. Epigenetic gene regulation in cancer. Adv Genet 61: 247-267. doi: 10.1016/S0065-2660(07)00009-0

Barolo S. 2012. Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. Bioessays 34: 135-141. doi: 10.1002/bies.201100121

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340-345. doi: Doi 10.1038/Ng.78

Barrios-Rodiles M, Brown KR, Ozdamar B, Bose R, Liu Z, Donovan RS, Shinjo F, Liu Y, Dembowy J, Taylor IW, Luga V, Przulj N, Robinson M, Suzuki H, Hayashizaki Y, Jurisica I, Wrana JL. 2005. High-throughput mapping of a dynamic signaling network in mammalian cells. Science 307: 1621-1625. doi: 10.1126/science.1105776

Bedford T, Hartl DL. 2008. Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in Drosophila. Mol Biol Evol 25: 1631-1638. doi: 10.1093/molbev/msn112

Benos PV, Bulyk ML, Stormo GD. 2002. Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res 30: 4442-4451.

Berg J, Willmann S, Lassig M. 2004a. Adaptive evolution of transcription factor binding sites. BMC Evol Biol 4: 42. doi: 10.1186/1471-2148-4-42

Berg J, Willmann S, Lässig M. 2004b. Adaptive evolution of transcription factor binding sites. BMC Evol Biol 4: 42. doi: 10.1186/1471-2148-4-42

Bergman CM, Carlson JW, Celniker SE. 2005. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics 21: 1747-1749.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A 99: 757-762. doi: 10.1073/pnas.231608898

Bradley RK, Holmes I. 2009. Evolutionary triplet models of structured RNA. PLoS Comput Biol 5: e1000483. doi: 10.1371/journal.pcbi.1000483

Brown CD, Johnson DS, Sidow A. 2007. Functional architecture and evolution of transcriptional elements that drive gene coexpression. Science 317: 1557-1560. doi: 317/5844/1557 [pii]

10.1126/science.1145893

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36: D102-106. doi: 10.1093/nar/gkm955

Buchler NE, Gerland U, Hwa T. 2003. On schemes of combinatorial transcription logic. Proc Natl Acad Sci U S A 100: 5136-5141. doi: 10.1073/pnas.0930314100

Burz DS, Rivera-Pomar R, Jackle H, Hanes SD. 1998. Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the Drosophila embryo. EMBO J 17: 5998-6009. doi: 10.1093/emboj/17.20.5998

Carter AJ, Wagner GP. 2002. Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. Proc Biol Sci 269: 953-960. doi: 10.1098/rspb.2002.1968

Cheng Q, Kazemian M, Pham H, Blatti C, Celniker SE, Wolfe SA, Brodsky MH, Sinha S. 2013. Computational Identification of Diverse Mechanisms Underlying Transcription Factor-DNA Occupancy. PLoS Genet 9: e1003571. doi: 10.1371/journal.pgen.1003571

Coleman RA, Pugh BF. 1995. Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA. J Biol Chem 270: 13850-13859.

Cooper MB, Loose M, Brookfield JF. 2009. The evolutionary influence of binding site organisation on gene regulatory networks. Biosystems 96: 185-193. doi: 10.1016/j.biosystems.2009.02.001

Damjanovski S, Huynh MH, Motamed K, Sage EH, Ringuette M. 1998. Regulation of SPARC expression during early Xenopus development: evolutionary divergence and conservation of DNA regulatory elements between amphibians and mammals. Dev Genes Evol 207: 453-461.

Davidson EH. 2010. The regulatory genome: gene regulatory networks in development and evolution: Academic Press.

de Souza FS, Franchini LF, Rubinstein M. 2013. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? Mol Biol Evol 30: 1239-1251. doi: 10.1093/molbev/mst045

Dearolf CR, Topol J, Parker CS. 1989. The caudal gene product is a direct activator of fushi tarazu transcription during Drosophila embryogenesis. Nature 341: 340-343. doi: 10.1038/341340a0

Dermitzakis ET, Bergman CM, Clark AG. 2003. Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. Mol Biol Evol 20: 703-714. doi: 10.1093/molbev/msg077

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 19: 1114-1121.

Doniger SW, Fay JC. 2007. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol 3: e99. doi: 10.1371/journal.pcbi.0030099

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics 148: 1667-1686.

Duque T, Samee MAH, Kazemian M, Pham HN, Brodsky MH, Sinha S. 2013. Simulations of enhancer evolution provide mechanistic insights into gene regulation. Mol Biol Evol: mst170.

Durrett R, Schmidt D. 2007. Waiting for regulatory sequences to appear. Annals of Applied Probability 17: 1-32. doi: Doi 10.1214/10505160600000619

Durrett R, Schmidt D. 2008. Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution. Genetics 180: 1501-1509. doi: 10.1534/genetics.107.082610

Emberly E, Rajewsky N, Siggia ED. 2003. Conservation of regulatory elements between two species of Drosophila. BMC Bioinformatics 4: 57. doi: 10.1186/1471-2105-4-57

Emera D, Casola C, Lynch VJ, Wildman DE, Agnew D, Wagner GP. 2012. Convergent evolution of endometrial prolactin expression in primates, mice, and elephants through the independent recruitment of transposable elements. Mol Biol Evol 29: 239-247. doi: 10.1093/molbev/msr189

ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9: e1001046. doi: 10.1371/journal.pbio.1001046

Fakhouri WD, Ay A, Sayal R, Dresch J, Dayringer E, Arnosti DN. 2010. Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. Mol Syst Biol 6: 341. doi: 10.1038/msb.2009.97

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17: 368-376.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet 7: 85-97. doi: 10.1038/nrg1767

Fisher AG. 2002. Cellular identity and lineage choice. Nat Rev Immunol 2: 977-982. doi: 10.1038/nri958

Fisher RA. 1999. The genetical theory of natural selection: a complete variorum edition: Oxford University Press.

Fowlkes CC, Eckenrode KB, Bragdon MD, Meyer M, Wunderlich Z, Simirenko L, Luengo Hendriks CL, Keranen SV, Henriquez C, Knowles DW, Biggin MD, Eisen MB, DePace AH. 2011. A conserved developmental patterning network produces quantitatively different output in multiple species of Drosophila. PLoS Genet 7: e1002346. doi: 10.1371/journal.pgen.1002346

Francois P, Hakim V, Siggia ED. 2007. Deriving structure from evolution: metazoan segmentation. Mol Syst Biol 3: 154. doi: 10.1038/msb4100192

Furriols M, Casanova J. 2003. In and out of Torso RTK signalling. EMBO J 22: 1947-1952. doi: 10.1093/emboj/cdg224

Gerland U, Hwa T. 2002. On the selection and evolution of regulatory DNA motifs. J Mol Evol 55: 386-400. doi: 10.1007/s00239-002-2335-z

Gertz J, Siggia ED, Cohen BA. 2009. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature 457: 215-218. doi: nature07521 [pii]

10.1038/nature07521

Giniger E, Ptashne M. 1988. Cooperative DNA binding of the yeast transcriptional activator GAL4. Proc Natl Acad Sci U S A 85: 382-386.

Goldberg DE. 2002. The Design of Innovation: Lessons from and for Competent Genetic Algorithms: Springer.

Goldberg DE. 1989. Genetic algorithms in search, optimization and machine learning.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res 20: 565-577. doi: 10.1101/gr.104471.109

Gray S, Levine M. 1996. Transcriptional repression in development. Curr Opin Cell Biol 8: 358-364.

Halfon MS, Gallo SM, Bergman CM. 2008a. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. Nucleic Acids Res 36: D594-598. doi: 10.1093/nar/gkm876

Halfon MS, Gallo SM, Bergman CM. 2008b. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. Nucleic Acids Res 36: D594-598.

Hallikas O, Taipale J. 2006. High-throughput assay for determining specificity and affinity of protein-DNA binding interactions. Nat Protoc 1: 215-222. doi: 10.1038/nprot.2006.33

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15: 910-917.

Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. 2008. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet 4: e1000106. doi: 10.1371/journal.pgen.1000106

Harrison MM, Li XY, Kaplan T, Botchan MR, Eisen MB. 2011. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet 7: e1002266. doi: 10.1371/journal.pgen.1002266

Hartl DL, Clark AG. 1997. Principles of population genetics: Sinauer associates Sunderland.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160-174.

He BZ, Holloway AK, Maerkl SJ, Kreitman M. 2011. Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. PLoS Genet 7: e1002053. doi: 10.1371/journal.pgen.1002053

He X, Chen CC, Hong F, Fang F, Sinha S, Ng HH, Zhong S. 2009. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. PLoS One 4: e8155. doi: 10.1371/journal.pone.0008155

He X, Duque TS, Sinha S. 2012. Evolutionary origins of transcription factor binding site clusters. Mol Biol Evol 29: 1059-1070. doi: 10.1093/molbev/msr277

He X, Samee MA, Blatti C, Sinha S. 2010. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput Biol 6. doi: 10.1371/journal.pcbi.1000935

Hedrick PW. 2011. Genetics of populations. Sudbury, Mass.: Jones and Bartlett Publishers.

Hein J, Schierup M, Wiuf C. 2004. Gene genealogies, variation and evolution: a primer in coalescent theory: Oxford university press.

Hertel KJ, Lynch KW, Maniatis T. 1997. Common themes in the function of transcription and splicing enhancers. Curr Opin Cell Biol 9: 350-357.

Hoekstra HE. 2006. Genetics, development and evolution of adaptive pigmentation in vertebrates. Heredity (Edinb) 97: 222-234. doi: 10.1038/sj.hdy.6800861

Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ. 2007. Sequence-level population simulations over large genomic regions. Genetics 177: 1725-1731. doi: 10.1534/genetics.106.069088

Holland JH. 1975. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence: U Michigan Press.

Holloway DM, Lopes FJ, da Fontoura Costa L, Travencolo BA, Golyandina N, Usevich K, Spirov AV. 2011. Gene expression noise in spatial patterning: hunchback promoter structure affects noise amplitude and distribution in Drosophila Segmentation. PLoS Comput Biol 7: e1001069. doi: 10.1371/journal.pcbi.1001069

Hordijk W. 1995. A Measure of Landscapes. Evolutionary Computation 4: 335--360.

Iwasa Y, Michor F, Nowak MA. 2004. Stochastic tunnels in evolutionary dynamics. Genetics 166: 1571-1579.

Jaeger J, Surkova S, Blagov M, Janssens H, Kosman D, Kozlov KN, Manu, Myasnikova E, Vanario-Alonso CE, Samsonova M, Sharp DH, Reinitz J. 2004. Dynamic control of positional information in the early Drosophila embryo. Nature 430: 368-371. doi: 10.1038/nature02678

Janssens H, Hou S, Jaeger J, Kim AR, Myasnikova E, Sharp D, Reinitz J. 2006. Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. Nat Genet 38: 1159-1165. doi: 10.1038/ng1886

Johnson RA, Wichern DW, Education P. 1992. Applied multivariate statistical analysis: Prentice hall Englewood Cliffs, NJ.

Josephides C, Moses AM. 2011. Modeling the evolution of a classic genetic switch. BMC Syst Biol 5: 24. doi: 10.1186/1752-0509-5-24

Joung JK, Le LU, Hochschild A. 1993. Synergistic activation of transcription by Escherichia coli cAMP receptor protein. Proc Natl Acad Sci U S A 90: 3083-3087.

Kaern M, Elston TC, Blake WJ, Collins JJ. 2005. Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet 6: 451-464. doi: nrg1615 [pii]

10.1038/nrg1615

Kanodia JS, Liang HL, Kim Y, Lim B, Zhan M, Lu H, Rushlow CA, Shvartsman SY. 2012. Pattern formation by graded and uniform signals in the early Drosophila embryo. Biophys J 102: 427-433. doi: 10.1016/j.bpj.2011.12.042

Kaplan T, Li XY, Sabo PJ, Thomas S, Stamatoyannopoulos JA, Biggin MD, Eisen MB. 2011. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genet 7: e1001290. doi: 10.1371/journal.pgen.1001290

Kazemian M, Blatti C, Richards A, McCutchan M, Wakabayashi-Ito N, Hammonds AS, Celniker SE, Kumar S, Wolfe SA, Brodsky MH, Sinha S. 2010. Quantitative analysis of the Drosophila segmentation regulatory network using pattern generating potentials. PLoS Biol 8. doi: 10.1371/journal.pbio.1000456

Kazemian M, Pham H, Wolfe SA, Brodsky MH, Sinha S. 2013. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. Nucleic Acids Res. doi: 10.1093/nar/gkt598

Khatri BS, McLeish TC, Sear RP. 2009. Statistical mechanics of convergent evolution in spatial patterning. Proc Natl Acad Sci U S A 106: 9564-9569. doi: 0812260106 [pii]

10.1073/pnas.0812260106

Kim AR, Martinez C, Ionides J, Ramos AF, Ludwig MZ, Ogawa N, Sharp DH, Reinitz J. 2013. Rearrangements of 2.5 kilobases of noncoding DNA from the Drosophila even-skipped locus define predictive rules of genomic cis-regulatory logic. PLoS Genet 9: e1003243. doi: 10.1371/journal.pgen.1003243

Kim J, He X, Sinha S. 2009. Evolution of regulatory sequences in 12 Drosophila species. PLoS Genet 5: e1000330. doi: 10.1371/journal.pgen.1000330

Kim J, Sinha S. 2010. Towards realistic benchmarks for multiple alignments of non-coding sequences. BMC Bioinformatics 11: 54. doi: 10.1186/1471-2105-11-54

Kim JG, Takeda Y, Matthews BW, Anderson WF. 1987. Kinetic studies on Cro repressor-operator DNA interaction. J Mol Biol 196: 149-158.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111-120.

Kimura M, Ohta T. 1969. The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. Genetics 61: 763-771.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M. 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420-426. doi: 10.1126/science.1149504

Lebrecht D, Foehr M, Smith E, Lopes FJ, Vanario-Alonso CE, Reinitz J, Burz DS, Hanes SD. 2005. Bicoid cooperative DNA binding is critical for embryonic patterning in Drosophila. Proc Natl Acad Sci U S A 102: 13176-13181. doi: 10.1073/pnas.0506462102

Levine M, Davidson EH. 2005. Gene regulatory networks for development. Proc Natl Acad Sci U S A 102: 4936-4942. doi: 10.1073/pnas.0408031102

Li G, Qian H. 2003. Sensitivity and specificity amplification in signal transduction. Cell Biochem Biophys 39: 45-59. doi: 10.1385/CBB:39:1:45

Li L, Zhu Q, He X, Sinha S, Halfon M. 2007a. Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. Genome Biol 8: R101.

Li L, Zhu Q, He X, Sinha S, Halfon MS. 2007b. Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. Genome Biol 8: R101. doi: gb-2007-8-6-r101 [pii]

10.1186/gb-2007-8-6-r101

Liang HL, Nien CY, Liu HY, Metzstein MM, Kirov N, Rushlow C. 2008. The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nature 456: 400-403. doi: 10.1038/nature07388

Liberman LM, Stathopoulos A. 2009. Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence. Dev Biol 327: 578-589. doi: 10.1016/j.ydbio.2008.12.020

Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA. 2003. Homotypic regulatory clusters in Drosophila. Genome Res 13: 579-588. doi: 10.1101/gr.668403

Lin YS, Carey M, Ptashne M, Green MR. 1990. How different eukaryotic transcriptional activators can cooperate promiscuously. Nature 345: 359-361. doi: 10.1038/345359a0

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

Ludwig MZ, Patel NH, Kreitman M. 1998. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development 125: 949-958.

Lusk RW, Eisen MB. 2010. Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. PLoS Genet 6: e1000829. doi: 10.1371/journal.pgen.1000829

Lynch M. 2007a. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A 104 Suppl 1: 8597-8604. doi: 0702207104 [pii]

10.1073/pnas.0702207104

Lynch M. 2007b. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A 104 Suppl 1: 8597-8604. doi: 10.1073/pnas.0702207104

MacArthur S, Brookfield JF. 2004. Expected rates and modes of evolution of enhancer sequences. Mol Biol Evol 21: 1064-1073. doi: 10.1093/molbev/msh105

Markstein M, Markstein P, Markstein V, Levine MS. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc Natl Acad Sci U S A 99: 763-768. doi: 10.1073/pnas.012591199

Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, Wilson RJ. 2013. FlyBase: improvements to the bibliography. Nucleic Acids Res 41: D751-757. doi: 10.1093/nar/gks1024

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW. 2013. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. doi: 10.1093/nar/gkt997

Meinhardt H. 1978. Space-dependent cell determination under the control of morphogen gradient. J Theor Biol 74: 307-321.

Moses AM. 2009. Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. BMC Evol Biol 9: 286. doi: 10.1186/1471-2148-9-286

Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB. 2004. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol 5: R98. doi: 10.1186/gb-2004-5-12-r98

Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB. 2006. Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol 2: e130. doi: 10.1371/journal.pcbi.0020130

Mustonen V, Kinney J, Callan CG, Jr., Lassig M. 2008. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proc Natl Acad Sci U S A 105: 12376-12381. doi: 10.1073/pnas.0805909105

Mustonen V, Lässig M. 2005. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc Natl Acad Sci U S A 102: 15936-15941. doi: 10.1073/pnas.0505537102

Neher RA. 2013. Genetic draft, selective interference, and population genetics of rapid adaptation. arXiv preprint arXiv:1302.1148.

Nien CY, Liang HL, Butcher S, Sun Y, Fu S, Gocha T, Kirov N, Manak JR, Rushlow C. 2011. Temporal coordination of gene networks by Zelda in the early Drosophila embryo. PLoS Genet 7: e1002339. doi: 10.1371/journal.pgen.1002339

Nourmohammad A, Lässig M. 2011. Formation of regulatory modules by local sequence duplication. PLoS Comput Biol 7: e1002167. doi: 10.1371/journal.pcbi.1002167

Noyes MB, Meng XD, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA. 2008. A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Res 36: 2547-2560. doi: Doi 10.1093/Nar/Gkn048

Ochoa-Espinosa A, Yucel G, Kaplan L, Pare A, Pura N, Oberstein A, Papatsenko D, Small S. 2005. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. Proc Natl Acad Sci U S A 102: 4960-4965. doi: 10.1073/pnas.0500373102

Okada N, Sasaki T, Shimogori T, Nishihara H. 2010. Emergence of mammals by emergency: exaptation. Genes Cells 15: 801-812. doi: 10.1111/j.1365-2443.2010.01429.x

Orr HA, Otto SP. 1994. Does diploidy increase the rate of adaptation? Genetics 136: 1475-1480.

Ostuni R, Piccolo V, Barozzi I, Polletti S, Termanini A, Bonifacio S, Curina A, Prosperini E, Ghisletti S, Natoli G. 2013. Latent enhancers activated by stimulation in differentiated cells. Cell 152: 157-171. doi: 10.1016/j.cell.2012.12.018

Paixao T, Azevedo RB. 2010. Redundancy and the evolution of cis-regulatory element multiplicity. PLoS Comput Biol 6: e1000848. doi: 10.1371/journal.pcbi.1000848

Papatsenko D, Goltsev Y, Levine M. 2009. Organization of developmental enhancers in the Drosophila embryo. Nucleic Acids Res 37: 5665-5677. doi: 10.1093/nar/gkp619

Papatsenko D, Levine MS. 2008. Dual regulation by the Hunchback gradient in the Drosophila embryo. Proc Natl Acad Sci U S A 105: 2901-2906. doi: 10.1073/pnas.0711941105

Parker DS, White MA, Ramos AI, Cohen BA, Barolo S. 2011. The cis-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. Sci Signal 4: ra38. doi: 10.1126/scisignal.2002077

Paten B, Herrero J, Beal K, Birney E. 2009. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. Bioinformatics 25: 295-301. doi: 10.1093/bioinformatics/btn630

Perkins TJ, Jaeger J, Reinitz J, Glass L. 2006. Reverse engineering the gap gene network of Drosophila melanogaster. PLoS Comput Biol 2: e51. doi: 10.1371/journal.pcbi.0020051

Perry MW, Boettiger AN, Bothma JP, Levine M. 2010. Shadow enhancers foster robustness of Drosophila gastrulation. Curr Biol 20: 1562-1567. doi: 10.1016/j.cub.2010.07.043

Perry MW, Boettiger AN, Levine M. 2011. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. Proc Natl Acad Sci U S A 108: 13570-13575. doi: 10.1073/pnas.1109873108

Porcher A, Dostatni N. 2010. The bicoid morphogen system. Curr Biol 20: R249-254. doi: 10.1016/j.cub.2010.01.026

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A 104 Suppl 1: 8605-8612. doi: 10.1073/pnas.0700488104

Pujato M, MacCarthy T, Fiser A, Bergman A. 2013. The underlying molecular and network level mechanisms in the evolution of robustness in gene regulatory networks. PLoS Comput Biol 9: e1002865. doi: 10.1371/journal.pcbi.1002865

Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science 300: 1742-1745. doi: 10.1126/science.1085881

Raser JM, O'Shea EK. 2004. Control of stochasticity in eukaryotic gene expression. Science 304: 1811-1814. doi: 10.1126/science.1098641

Reid ID. 2007. Transcription Factor Binding Site Turnover in Mammals. [[Montreal]: McGill University.

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, Couronne O, Hua S, Smith MA, Zhang P, Liu J, Bussemaker HJ, van Batenburg MF, Howells SL, Scherer SE, Sodergren E, Matthews BB, Crosby MA, Schroeder AJ, Ortiz-Barrientos D, Rives CM, Metzker ML, Muzny DM, Scott G, Steffen D, Wheeler DA, Worley KC, Havlak P, Durbin KJ, Egan A, Gill R, Hume J, Morgan MB, Miner G, Hamilton C, Huang Y, Waldron L, Verduzco D, Clerc-Blankenburg KP, Dubchak I, Noor MA, Anderson W, White KP, Clark AG, Schaeffer SW, Gelbart W, Weinstock GM, Gibbs RA. 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res 15: 1-18. doi: 10.1101/gr.3059305

Riggs AD, Jones PA. 1983. 5-methylcytosine, gene regulation, and cancer. Adv Cancer Res 40: 1-30.

Roider HG, Kanhere A, Manke T, Vingron M. 2007. Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics 23: 134-141. doi: btl565 [pii]

10.1093/bioinformatics/btl565

Ross JL, Fong PP, Cavener DR. 1994. Correlated evolution of the cis-acting regulatory elements and developmental expression of the Drosophila Gld gene in seven species from the subgroup melanogaster. Dev Genet 15: 38-50. doi: 10.1002/dvg.1020150106

Samee MA, Sinha S. 2013. Evaluating thermodynamic models of enhancer activity on cellular resolution gene expression data. Methods. doi: 10.1016/j.ymeth.2013.03.005

Satta Y, Ishiwa H, Chigusa SI. 1987. Analysis of nucleotide substitutions of mitochondrial DNAs in Drosophila melanogaster and its sibling species. Mol Biol Evol 4: 638-650.

Satta Y, Takahata N. 1990. Evolution of Drosophila mitochondrial DNA and the history of the melanogaster subgroup. Proc Natl Acad Sci U S A 87: 9558-9562.

Sauer F, Hansen SK, Tjian R. 1995. DNA template and activator-coactivator requirements for transcriptional synergism by Drosophila bicoid. Science 270: 1825-1828.

Schoustra SE, Debets AJ, Slakhorst M, Hoekstra RF. 2007. Mitotic recombination accelerates adaptation in the fungus Aspergillus nidulans. PLoS Genet 3: e68. doi: 10.1371/journal.pgen.0030068

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 451: 535-540. doi: 10.1038/nature06496

Segal JA, Barnett JL, Crawford DL. 1999. Functional analyses of natural variation in Sp1 binding sites of a TATA-less promoter. J Mol Evol 49: 736-749.

Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci U S A 102: 9541-9546. doi: 10.1073/pnas.0501865102

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

Shultzaberger RK, Malashock DS, Kirsch JF, Eisen MB. 2010. The fitness landscapes of cis-acting binding sites in different promoter and environmental contexts. PLoS Genet 6: e1001042. doi: 10.1371/journal.pgen.1001042

Sinha S, Adler AS, Field Y, Chang HY, Segal E. 2008. Systematic functional characterization of cis-regulatory motifs in human core promoters. Genome Res 18: 477-488. doi: 10.1101/gr.6828808

Sinha S, Liang Y, Siggia E. 2006. Stubb: a program for discovery and analysis of cis-regulatory modules. Nucleic Acids Res 34: W555-559. doi: 10.1093/nar/gkl224

Sinha S, Siggia ED. 2005. Sequence turnover and tandem repeats in cis-regulatory modules in drosophila. Mol Biol Evol 22: 874-885. doi: 10.1093/molbev/msi090

Sinha S, van Nimwegen E, Siggia ED. 2003. A probabilistic method to detect regulatory modules. Bioinformatics 19 Suppl 1: i292-301.

Smith T, Husbands P, Layzell P, M. OS. 2002. Fitness landscapes and evolvability. Evolutionary Computation 10: 1-34.

Soyer OS, Bonhoeffer S. 2006. Evolution of complexity in signaling pathways. Proc Natl Acad Sci U S A 103: 16337-16342. doi: 10.1073/pnas.0604449103

Spivakov M, Akhtar J, Kheradpour P, Beal K, Girardot C, Koscielny G, Herrero J, Kellis M, Furlong EE, Birney E. 2012. Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biol 13: R49. doi: 10.1186/gb-2012-13-9-r49

St Johnston D, Nusslein-Volhard C. 1992. The origin of pattern and polarity in the Drosophila embryo. Cell 68: 201-219.

Staten R, Schully SD, Noor MA. 2004. A microsatellite linkage map of Drosophila mojavensis. BMC Genet 5: 12. doi: 10.1186/1471-2156-5-12

Stewart AJ, Seymour RM, Pomiankowski A, Plotkin JB. 2012. The population genetics of cooperative gene regulation. BMC Evol Biol 12: 173. doi: 10.1186/1471-2148-12-173

Stone JR, Wray GA. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol 18: 1764-1770.

Stormo GD. 2000. DNA binding sites: representation and discovery. Bioinformatics 16: 16-23.

Stormo GD, Fields DS. 1998. Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci 23: 109-113.

Suh E, Chen L, Taylor J, Traber PG. 1994. A homeodomain protein related to caudal regulates intestine-specific gene transcription. Mol Cell Biol 14: 7340-7351.

Swanson CI, Schwimmer DB, Barolo S. 2011. Rapid evolutionary rewiring of a structurally constrained eye enhancer. Curr Biol 21: 1186-1196. doi: 10.1016/j.cub.2011.05.056

Tanay A, Siggia ED. 2008. Sequence context affects the rate of short insertions and deletions in flies and primates. Genome Biol 9: R37. doi: 10.1186/gb-2008-9-2-r37

Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics 172: 1607-1619. doi: 10.1534/genetics.105.048223

Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, Rubin GM. 2002. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol 3: RESEARCH0088.

Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM. 2007. Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol 8: R145. doi: 10.1186/gb-2007-8-7-r145

Tsurumi A, Xia F, Li J, Larson K, LaFrance R, Li WX. 2011. STAT is an essential activator of the zygotic genome in the early Drosophila embryo. PLoS Genet 7: e1002086. doi: 10.1371/journal.pgen.1002086

van Duijn CM, Cruts M, Theuns J, Van Gassen G, Backhovens H, van den Broeck M, Wehnert A, Serneels S, Hofman A, Van Broeckhoven C. 1999. Genetic association of the presenilin-1 regulatory region with early-onset Alzheimer's disease in a population-based sample. Eur J Hum Genet 7: 801-806. doi: 10.1038/sj.ejhg.5200373

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet 40: 158-160. doi: 10.1038/ng.2007.55

Vizoso Pinto MG, Villegas JM, Peter J, Haase R, Haas J, Lotz AS, Muntau AC, Baiker A. 2009. LuMPIS--a modified luminescence-based mammalian interactome mapping pull-down assay for the investigation of protein-protein interactions encoded by GC-low ORFs. Proteomics 9: 5303-5308. doi: 10.1002/pmic.200900298

Wagner A. 2007. Robustness and Evolvability in Living Systems: Princeton University Press.

Wagner A. 2005. Robustness and evolvability in living systems: Princeton University Press Princeton.

Wagner GP, Altenberg L. 1996. Perspective: Complex adaptations and the evolution of evolvability. Evolution 50: 967-976. doi: Doi 10.2307/2410639

Weinberger ED. 1991. Local properties of Kauffman's N-k model: A tunably rugged energy landscape. Phys. Rev. A 44: 6399–6413.

Weirauch MT, Hughes TR. 2010. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet 26: 66-74. doi: 10.1016/j.tig.2009.12.002

Whittaker J. 2009. Graphical models in applied multivariate statistics: Wiley Publishing.

Wingender E. 2008. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief Bioinform 9: 326-332. doi: 10.1093/bib/bbn016

Wright S. 1931. Evolution in Mendelian Populations. Genetics 16: 97-159.

Yao LC, Phin S, Cho J, Rushlow C, Arora K, Warrior R. 2008. Multiple modular promoter elements drive graded brinker expression in response to the Dpp morphogen gradient. Development 135: 2183-2192. doi: 10.1242/dev.015826

Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, Young RA, Levine M. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev 21: 385-390. doi: 10.1101/gad.1509607

Zeyl C, Vanderford T, Carter M. 2003. An evolutionary advantage of haploidy in large yeast populations. Science 299: 555-558. doi: 10.1126/science.1078417

Zinzen RP, Senger K, Levine M, Papatsenko D. 2006. Computational models for neurogenic gene expression in the Drosophila embryo. Curr Biol 16: 1358-1365. doi: 10.1016/j.cub.2006.05.044