

Studies into Location-specific cis-Regulatory Motifs

by

Ken Daigoro Yokoyama

Department of Biology
Duke University

Date: _____

Approved:

Lindsay G. Cowell, Committee Chair

Gregory A. Wray, Advisor

Terry Furey

Sayan Mukherjee

Uwe Ohler

Jeffrey L. Thorne

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Biology
in the Graduate School of Duke University
2010

ABSTRACT
(Bioinformatics)

Studies into Location-specific cis-Regulatory Motifs

by

Ken Daigoro Yokoyama

Department of Biology
Duke University

Date: _____

Approved:

Lindsay G. Cowell, Committee Chair

Gregory A. Wray, Advisor

Terry Furey

Sayan Mukherjee

Uwe Ohler

Jeffrey L. Thorne

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Biology
in the Graduate School of Duke University

2010

Copyright © 2010 by Ken Daigoro Yokoyama
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

Gene expression and regulation are major determinants of phenotypic traits displayed across species. Although the DNA sequence elements that control gene expression play a crucial role in determining species morphology, predicting *cis*-regulatory elements through sequence analysis alone remains a difficult task. A few regulatory elements, such as the TATA-box and Initiator sequence, are overrepresented at specific locations within the proximal promoter. However, the extent to which this occurs among *cis*-regulatory elements is not well understood. Here, we take a genome-wide approach towards detecting such functional sequence elements, using location-specific overrepresentation as a criterion for regulatory function. We provide evidence that a surprisingly large number of regulatory elements exhibit locational overrepresentation with respect to the transcription start site. We then utilize this characteristic to predict novel *cis*-regulatory elements overrepresented at particular locations within the proximal promoter.

Transcriptional regulation is most often controlled not by single protein factors acting in isolation, but instead multiple transcription factors acting together within multi-protein complexes. As protein-protein interactions are largely determined through protein structure, we would expect to see patterns of spatial preference between motif-pairs binding interacting factors. However, in the absence of methods to predict such spatial preferences between motifs, comprehensive assessments of such inter-relationships have not been previously conducted. As our model provides a general tool for detecting positional specificities of a motif relative to a given reference point, we expanded our model to measure distance preferences between pairs of motifs on a genome-wide scale. We show that there often exist patterns of spatial dependencies between pairs of sequence elements that bind interacting protein factors. We find that regulatory motifs binding interacting proteins often have multiple inter-motif distances at which they preferentially occur, and we show that the intervals between preferred distances are highly consistent across motif-pairs. This distance preference ‘phasing’ was empirically found to occur at consistent intervals

around ~ 8 -10 bp, corresponding to approximately the number of nucleotides within a single turn of the DNA double-helix. This finding suggests a tendency for protein factor-pairs to interact in a specific orientation with respect to the turn of the DNA molecule, and offers a convenient method by which to determine motif-pairs binding interacting transcription factors *de novo*.

While little is known about the mechanisms by which individual *cis*-regulatory elements ultimately control gene expression, even less is known about how such elements evolve over time. A single transcription factor can potentially target hundreds of genes across the genome, and thus modifications in the binding affinities of such proteins must induce conversions at a multitude of functional sites in order to preserve the set of target genes that the *trans*-factor regulates. It is therefore commonly assumed that such changes occur rarely and at a slow rate over the course of evolution. Despite this widespread assumption, we find that a surprisingly large number of *cis*-regulatory elements have been subject to significant changes in consensus sequence in a lineage-specific manner. Here, we demonstrate that the genomic landscape is highly adaptable, rapidly adjusting to global changes in preferred regulatory consensus sequences. Focusing upon regulatory elements exhibiting location-specific overrepresentation, we find that a substantial fraction of regulatory elements have been subject to evolutionary modifications, even between closely related eutherians. These findings have broad implications regarding evolving phenotypes observed across species.

To all of those who have taught and inspired me throughout my life.
Thank you all.

Contents

List of Tables	xii
List of Figures	xiii
List of Abbreviations and Symbols	xiv
Acknowledgements	xv
1 Introduction	1
1.1 Regulatory motif prediction and locational specificity	2
1.2 Spatial patterns of functional regulatory motif inter-dependencies . .	4
1.3 Evolution of location-specific <i>cis</i> -regulatory elements	6
2 Background	10
2.1 Transcriptional gene regulation	10
2.1.1 Gene regulation via transcription factors	10
2.1.2 Alternate mechanisms of transcriptional regulation	12
2.1.3 Regulatory sequence analysis	13
2.2 Available genomic sequence data	14
2.2.1 RefSeq gene annotations	14
2.2.2 Genome projects across diverse vertebrate species	15
2.3 Locational specificity in <i>cis</i> -regulatory elements	16
2.3.1 Promoter architecture	16
2.3.2 Locational-specificity within regulatory motifs	18

2.3.3	Assessments of locational specificity	20
2.4	Mutual relationships between regulatory elements	21
2.4.1	Combinatorial regulatory element relationships	21
2.4.2	Distance relationships between regulatory motifs	22
2.4.3	Periodic distributions of inter-motif distance preferences	23
2.5	Evolution of transcriptional regulation	24
2.5.1	Significance of gene expression on the evolving organism	24
2.5.2	Preservation of gene expression patterns versus sequence conservation	25
2.5.3	Evolution of cis-regulatory element consensus sequences	26
3	Motif Locational Functions	29
3.1	Motif locational functions (MLFs) predict locational specificity in <i>cis</i> -regulatory motifs	30
3.1.1	Overview of the MLF model	30
3.1.2	Accounting for dinucleotide fluctuations within the promoter	31
3.2	The MLF method predicts location-specific overrepresentation for many motifs within human promoters	34
3.2.1	Predicting locational overrepresentation according to results from control data sets	34
3.2.2	Motif clustering predicts locational overrepresentation for both known and putatively novel <i>cis</i> -regulatory elements	36
3.3	Many location-specific motifs are shared between human and mouse	39
3.4	Study comparisons highlight differences in methodologies to previous studies	41
3.4.1	The ‘sliding window’ approach	41
3.4.2	Comparisons to Tharakaraman et al.	43
3.4.3	Comparisons to Vardhanabhati et al.	45

4	Motif Relational Functions	48
4.1	Motif relational functions (MRFs) detect spatial biases between motif-pairs	49
4.1.1	Overview of the MRF model	49
4.1.2	The MRF model incorporates phased distance preferences between motifs	50
4.1.3	Phased intervals between preferred inter-motif distances corresponds to the turn of the DNA double-helix	52
4.2	Predicting regulatory motifs using peak separation values	53
4.2.1	Several motifs exhibit consistent peak separation values	53
4.2.2	Periodic phasing of inter-motif distance preferences detects known and novel regulatory element relationships	56
4.3	Uni-modal versus multi-modal approaches	58
4.4	Deviation of phasing interval values	59
5	Location-specific <i>cis</i>-regulatory element evolution	61
5.1	Studying location-specific <i>cis</i> -regulatory element evolution	63
5.1.1	Location-specific motifs are shared across closely and distantly related species	63
5.1.2	Determining functional motif co-occurrences across species	64
5.1.3	Determining background motif co-occurrences	65
5.1.4	Motif conservation is stronger within the region of overrepresentation	66
5.2	Modeling lineage-specific regulatory motif modifications	67
5.2.1	Determining motif modifications according to nucleotide co-occurrence asymmetries	67
5.2.2	Intergenic background model	69
5.2.3	The binomial distribution model	70
5.3	Evolutionary modifications within location-specific motifs within vertebrates	71

5.3.1	Genome-wide biases in nucleotide preferences are stronger within the region of overrepresentation	71
5.3.2	Pair-wise species comparisons conducted within vertebrates . .	73
5.3.3	Many <i>cis</i> -regulatory elements have undergone rapid evolutionary changes within mammals	74
5.3.4	Correlation between divergence time and motif modification .	76
5.3.5	Simulation analyses and the effect of multiple hypothesis testing	76
5.3.6	The high prevalence of motif modification is supported by the binomial distribution model	78
5.3.7	Both site degeneracy and preferred consensus nucleotides change over the course of evolution	79
5.3.8	The GC box regulatory motif has been modified along the eutherian branch	80
5.4	Discussion	82
5.4.1	Use of location-specific regulatory motifs	83
5.4.2	Mechanisms of <i>cis</i> -regulatory element modification	84
6	Conclusions	86
6.1	Summary	86
6.2	Future directions	89
6.3	Concluding Remarks	92
A	MLF methodology	94
A.1	Statistical framework of the MLF model	94
A.2	Background functions	95
A.3	Parameter estimation and statistical significance determination	97
A.4	Model selection	98
A.5	Motif clustering procedure	99
A.6	Consensus sequence determination and known <i>cis</i> -regulatory motif comparisons	100

B MRF methodology	101
B.1 Statistical framework of the MRF model	101
B.2 Background functions	102
B.3 MRF parameter estimation	103
B.4 MRF model selection	103
C Derivation of the intergenic background model Z-score function	105
Bibliography	107
Biography	121

List of Tables

3.1	Location-specific motifs in human promoters	38
3.2	Between-studies comparisons of locational specificity predictions	42
3.3	‘Novel’ motif predictions by Vardhanabhuti et al	46
4.1	Consensus motifs with consistent phasing intervals	54
4.2	Motif partners for the NFY binding element	57
4.3	MADS-box family binding site partners	57
5.1	Location-specific regulatory motifs within mouse and zebrafish	64
5.2	Data sets used for cross-species comparisons	73
5.3	Prevalence of evolutionary modifications in regulatory motifs within mammals	75
5.4	Prevalence of modified motifs: Binomial distribution approach	79
5.5	Ten regulatory motifs differing between human and mouse	80
5.6	Evolutionary changes within the GC box element	81

List of Figures

2.1	The transcriptional complex	11
3.1	Motif locational functions (MLFs)	32
3.2	Histogram of p-value predictions from the MLF analysis	35
3.3	Location of overrepresentation for the 48 MLF motif predictions within human promoters	39
3.4	MLFs of six motifs exhibiting location-specific overrepresentation	40
3.5	Occurrence frequency of the functional Inr sequence	44
4.1	Functional motif-pair inter-relationships	50
4.2	Motif relational function (MRF) examples	51
4.3	Distribution of MRF peak separation values	53
4.4	MRF peak separation concentrations	55
5.1	Examples of <i>cis</i> -regulatory motif modifications	62
5.2	Nucleotide co-occurrence data for the NFY binding element	65
5.3	Nucleotide conservation frequencies according to location	67
5.4	Regulatory motif modification scheme	68
5.5	Cross-species asymmetries in nucleotide frequencies	72
5.6	Prevalence of evolutionary modifications versus divergence time	77

List of Abbreviations and Symbols

CRM	Cis-regulatory module.
GTF	General transcription factor.
MLF	Motif locational function.
MPF	Motif positional function.
MRF	Motif relational function.
TFBS	Transcription factor binding site.
TSS	Transcription start site.

Acknowledgements

I would like to thank my advisor Gregory A. Wray for giving his kind support throughout my graduate career. I would also like to thank both Uwe Ohler and Jeffrey L. Thorne for their hard work regarding my manuscripts and publications; without their help, this work would not have been possible. I would also like to acknowledge the rest of my committee for their help: Sayan Mukherjee, Lindsay Cowell, and Terry Furey. Many thanks to David Garfield, Courtney Babbitt, and Lisa Warner for their helpful input regarding my second manuscript, as well as the rest of the Wray lab and the Ohler lab. I would also like to thank my parents for their help, advice, and support. Thanks to all.

Introduction

Transcriptional initiation is a major point of control for gene expression [66, 94, 144], and considerable effort has been devoted to deciphering the code by which transcriptional regulation occurs. Although this aspect of the genotype-phenotype connection is central to many fundamental biological processes, our understanding of how this mechanism operates at the molecular level is far from complete.

Transcription is largely driven by proteins called transcription factors that bind to the DNA in a sequence-specific manner [66, 144]. The mechanisms by which transcription occurs are highly complex, and transcriptional regulation is often driven by multiple regulatory elements acting together to control gene expression [4, 10, 33, 35, 144]. However, despite such complexities, the rapidly-growing availability of genomic data has allowed for computational approaches to be used to study gene regulation at the DNA sequence level. While it is unlikely that sequence analysis alone will enable us to understand the entire range of biological processes within the organism, it serves as a major step towards our knowledge of how such processes occur.

In this work, we utilize sequence data on a genome-wide scale in order to un-

derstand various aspects of transcriptional regulation. Our approach is to assess spatial patterns of sequence element occurrence, such as locational specificity of DNA sequence motifs relative to the transcription start site (TSS) or positional inter-dependencies between pairs of sequence elements. In Chapters 3 and 4, we show that spatial preferences can be used both to discover *cis*-regulatory elements as well as functional relationships between regulatory motifs. Our results include many known *cis*-regulatory motifs [80] as well as a large number of novel regulatory motif candidates. We further show that spatial inter-dependencies exist between many motif-pairs binding interacting transcription factors, and we use this characteristic to predict previously undocumented relationships between putative regulatory motifs. In Chapter 5, we assess the prevalence of evolutionary changes within these location-specific elements across a large array of vertebrate species. We show that locational specificity can be utilized to study the evolution of such protein binding elements. Our studies represent a novel means by which to predict *cis*-regulatory motifs and their significance within the evolving organism.

1.1 Regulatory motif prediction and locational specificity

At the most fundamental level, understanding transcriptional regulation requires knowledge about the individual *cis*-regulatory elements that affect gene expression. Several methods have been proposed to predict individual transcription factor binding sites by detecting statistically overrepresented motifs within the promoter [3, 32, 62, 67, 93, 105, 121, 126, 127]. This approach generally searches for sequence elements common across a small handful of co-regulated genes, making the assumption that co-regulation implies co-occurrence of the same putative regulatory element. However, in the absence of functional data, overrepresentation of DNA sequence elements is not a sufficient criterion for functionality for two basic reasons. First, binding sites are generally short (~ 5 -10 bp) [124], which means that many

instances are present by chance rather than for functional reasons. Second, many motifs occur at increased frequency as a result of mutational bias or dinucleotide fluctuations near the start of transcription [38, 119, 149]. The widespread overrepresentation of these motifs frequently dominates the subtle indicators of regulatory function, thus limiting the efficacy of these approaches.

In this study, we take a genome-wide approach to predict *cis*-regulatory motifs, using locational specificity as a criterion for motif functionality. There are a few known cases for which *cis*-regulatory elements function at specific locations within the proximal promoter [38, 66, 79, 145]. Such regulatory elements are often overrepresented at the location at which they function with respect to the surrounding region [38, 128, 145]. This characteristic can therefore be utilized to determine functional *cis*-regulatory elements *de novo*.

Previously, relatively little information existed about locational specificity of *cis*-regulatory sequences. Important questions include whether location-specific regulatory motifs are common, how they are distributed around genes, which transcription factors are involved, and whether they are associated with functional classes of genes. Here, we report that a large number of known and novel regulatory elements exhibit location-specific overrepresentation relative to the TSS. We introduce a new model, denoted as the ‘motif positional function’ (MPF) model, which predicts regulatory elements by measuring locational specificity at single basepair resolution while considering the data collectively. By using regression analysis and a likelihood ratio test, the model differentiates noise in the data from true location-specific overrepresentation. The method also accounts for location-specific dinucleotide fluctuations that exist within the promoter, incorporating a non-uniformly distributed background (null) model based upon the dinucleotide composition across the regulatory region. We show that this method can be used to predict novel *cis*-regulatory elements exhibiting previously unrecognized instances of positional specificity on a genome-wide

scale.

1.2 Spatial patterns of functional regulatory motif inter-dependencies

Beyond a fundamental understanding of how individual *cis*-regulatory motifs affect gene expression, it is also crucial to understand how different regulatory elements work in concert to control gene expression. Transcription is not only regulated by individual proteins working in isolation, but is instead driven by cooperative interactions between multiple protein factors [4, 10, 33, 144]. Thus, knowledge regarding such regulatory element relationships is essential to gain a full understanding of gene regulation.

Since protein-protein interactions are inherently structure-specific, it is natural to expect that regulatory motifs binding interacting proteins preferentially co-occur non-randomly with respect to each other. Previous studies have shown that mutual relationships exist between various regulatory motifs, such as paired co-occurrences and relative orientations to the TSS [59, 70, 95, 112]. Such relationships have been effectively utilized in a variety of applications, such as the study of condition-specific and time-dependent gene expression patterns [5], gene network analyses [95, 118], and promoter region detection [75]. However, such studies are frequently limited to analyzing sequence element relationships between either known binding site motifs or those predicted using standard motif overrepresentation methods. Due to this limitation, little is known about the nature of *cis*-regulatory elements relationships, and tools designed for large-scale assessments of such relationships have not been previously available.

Here, we provide a means by which to predict pairs of regulatory motifs binding interacting transcription factors *de novo*, without any prior knowledge about the sequences of the particular motifs involved. Our model provides a general tool to measure positional specificity relative to any given reference point, and is therefore

not restricted to measure only location-specific preferences relative to the TSS. Thus, we extend our model to measure positional specificity of a given motif relative to the known occurrence of a second motif, analyzing spatial relationships between pairs of motifs with a collective functional role in gene regulation. Our approach effectively circumvents the limitations of previous studies, for which relationships could only be analyzed for previously known or predicted regulatory motifs.

It has been previously shown that protein-protein interactions can occur at phased intervals along the DNA molecule due to the winding of the double-helix or the wrapping within the histone complex [53, 69, 135, 144]. Although this has been shown in only a few individual cases, without large-scale assessments, the prevalence of such ‘interaction phasing’ has not been studied. In this work, we search comprehensively across all motif-pairs to discover potential binding site relationships, incorporating a multi-modal model for inter-motif distance preferences. Here, we show that interaction phasing is a very common characteristic among many motif-pairs, and we find that binding sites of putatively interacting transcription factors frequently co-occur preferentially at multiple distances. These preferred inter-motif distances were found to be consistently phased, with intervals between preferred distances corresponding approximately to the number of nucleotides in one turn of the DNA double-helix. This suggests a tendency for certain factor-pair interactions to occur in a particular orientation relative to the turn of the DNA molecule. We use this periodic phasing of inter-motif distance preferences to detect motif-pairs binding interacting proteins, predicting functional binding site relationships between putative regulatory motifs. We show that this methodology is effective in predicting pairs of regulatory motifs that bind interacting transcription factor proteins, including both individual regulatory motifs as well as motif-pairs binding proteins with known and previously undocumented interactions.

1.3 Evolution of location-specific *cis*-regulatory elements

Transcriptional regulation is a major determinant of species morphology [49, 66, 143, 144], and therefore evolutionary changes in *cis*-regulatory elements can have broad impacts on phenotypic traits observed across organisms [64, 122, 130, 143]. A single transcription factor can sometimes regulate the expression of hundreds or even thousands of genes genome-wide, particularly in the case of commonly-occurring location-specific motifs [38, 148]. Modifications within the preferred protein-binding sequences therefore involve a massive number of changes in order to preserve the set of target genes regulated by the corresponding *trans*-factor.

In this ‘one-to-many’ relationship between a single transcription factor and the multitude of sites to which it binds, it is often assumed that protein binding affinities, and thus the preferred DNA binding motifs, are rarely modified over the course of evolution [31, 44, 60, 61, 72, 128, 146]. This assumption is both convenient from a computational point of view and, in the absence of evidence to the contrary, may appear to be plausible. However, the assumption that regulatory motifs remain mostly static has never been explicitly tested. Large-scale analyses for cross-species differences in preferred binding motifs are typically unfeasible using experimental approaches, and in the absence of effective computational approaches, searching for such differences *in silico* has previously been proven difficult.

In this work, we take a novel approach to assess the prevalence and nature of genome-wide regulatory motif modifications within vertebrates. We focus upon motifs exhibiting location-specific overrepresentation as predicted using the MPF model. Locational specificity is a convenient characteristic by which to study evolutionary changes in *cis*-regulatory motifs. Location-specific overrepresentation reflects a functional role of a given motif within its region of overrepresentation [79, 149], and therefore we can infer functionality for motif occurrences at that location. This allows

us to focus upon instances of the motif that are likely to be functional, while other instances of the motif outside this region are less likely to play a role in gene regulation [128]. In contrast, distinguishing between functional occurrences and those under less evolutionary constraints is difficult for non-location-specific motifs. Many regulatory motif occurrences are simply due to chance and may not necessarily play a role in gene regulation, depending upon complex factors such as nearby sequence elements or structural characteristics of the surrounding DNA [45, 134, 144].

Although many occurrences of location-specific regulatory motifs are also present due to chance, previous observations have suggested that occurrences of location-specific motifs within their preferred locations are much more likely to be functional than occurrences outside this preferred location. For instance, it has been shown that occurrences within the region of overrepresentation are under stronger evolutionary constraints than those outside this region [128]. It has also been shown that several location-specific motifs induce similar expression patterns of the target genes when occurring within the region of overrepresentation [38], while other motifs target genes with specific GO categories [128]. These observations suggest that functionality of location-specific motifs can often be inferred according to the location of the occurrences, without the complexity of other factors affecting functionality of non-location-specific regulatory motifs.

In addition, locational specificity allows us to circumvent the need for cross-species alignments. In our approach, we simply consider motif occurrences targeting orthologous genes across species, focusing only upon occurrences within the region of overrepresentation. However, such occurrences need not be paired within cross-species alignments. This is a significant advantage over previous cross-species comparisons that rely upon sequence alignments [60, 61, 128, 141, 146], since obtaining reliable sequence alignments across highly diverged species is largely unfeasible. Thus, this approach enables us to assess regulatory element evolution across a wide

range of vertebrate species, including humans, chimp, macaque, mouse, rat, cow, dog, horse, opossum, platypus, lizard, and frog. In addition, without relying upon sequence alignments, our methodology naturally incorporates instances of regulatory element turnover within regions of overrepresentation, which has been shown to occur commonly during the course of evolution [44, 72, 91].

In this work, we conduct two different statistical assessments of the prevalence of regulatory motif modification, each according to a separate statistical model with differing underlying assumptions. The first, which we denote as the ‘intergenic background’ model, compares cross-species differences observed within the region of overrepresentation relative to the expected amount of non-conservation outside this region (i.e., within a set of intergenic sequences). The second, denoted as the ‘binomial distribution’ model, considers only motif occurrences within the region of overrepresentation, with the null model assuming an equal rate of substitution to and from each nucleotide type across lineages. With both models, we show that a surprisingly large fraction of location-specific regulatory elements have been subject to evolutionary modifications since the divergence between even relatively closely related species. For instance, we find that approximately a third of all location-specific motifs exhibit rapid evolutionary modifications within either human or mouse following species divergence. In many cases, regulatory motifs were modified at a rapid rate compared to the background frequency of substitution. In some cases, this genome-wide conversion reflects hundreds of nucleotide-specific substitutions each occurring independently across the genome. This was found to be the case for even well-studied regulatory elements such as the GC box, for which we found nucleotide-specific substitutions (gggAgg \rightarrow gggCgg) in approximately 600 promoters along the eutherian lineage following the split from marsupials.

These findings illustrate a highly adaptable nature of the genome. We show that the genome is able to accumulate a surprisingly high number of coordinated modi-

fications in functional sequence elements, with independently occurring nucleotide-specific substitutions. While many studies focus upon either a single or a small handful of genes to explain phenotypic differences across species, e.g., [31, 130], the genome-wide modifications discovered here affect a massive number of genes, and therefore have broad implications regarding phenotypic differences observed across species. These results provide motivation for future work regarding *cis*-regulatory element evolution.

2

Background

2.1 Transcriptional gene regulation

Since the landmark paper by King and Wilson in 1975 [64], it has been postulated that gene regulation is one of the major determinants of organismal phenotype. Of the many mechanisms by which gene regulation occurs, transcription is generally thought to be the major point of control [66, 144]. At the molecular level, transcription is the process by which the DNA template is used to create a complementary RNA molecule. RNA is then often, although not always, translated into a protein that performs a specific biological function. Below we describe the mechanisms by which the transcriptional process is controlled.

2.1.1 Gene regulation via transcription factors

Transcription is carried out by a multi-protein complex called the basal transcriptional complex [66, 142, 148], whose components include RNA polymerase and a large number of interacting proteins called ‘general transcription factors’ (GTFs) [142]. As shown in Figure 2.1, this complex binds to the DNA at a location called the core promoter and subsequently transcribes the DNA template into mRNA [66, 104, 142].

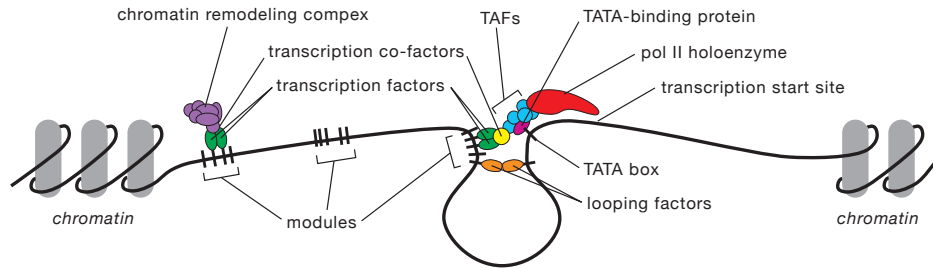


FIGURE 2.1: The transcriptional complex. Transcription is initiated via protein factor complexes that bind to DNA *cis*-regulatory elements. The basal transcriptional complex, comprised of transcription activating factors (TAFs) and RNA polymerase II, carries out transcription after binding to various promoter elements, such as the TATA-box. Transcription is also regulated by protein factors binding to regulatory DNA modules, often located farther from the transcription start site.

Although the binding of the basal transcriptional complex to the promoter is sufficient for minimal transcriptional initiation [104], additional protein factors are generally required to fully activate transcription [104, 142]. Such transcription factor proteins generally bind to the DNA in a sequence-specific manner, or else act in concert with other bound proteins via protein-protein interactions [144]. These proteins can either induce or inhibit transcription, according to the characteristics of the particular protein factor [66, 104].

The mechanisms by which transcription occurs are highly complex, and are largely driven through the collective efforts of multiple protein factors acting in concert [4, 10, 33, 144]. Although the expression of certain genes (e.g., housekeeping genes) remains relatively stable across different tissues and external conditions, in other genes, relationships between interacting factors can often direct complex patterns of expression. Such complex expression patterns are often associated with, for example, genes involved in cell differentiation and development [99] or signaling pathways [42]. Transcription factors controlling regulation often bind to the DNA either inside the proximal promoter region (within ~ 200 bp of the TSS), or in distal

enhancers, which can be located as far as 10 kb away of the start of transcription [66]. Binding sequences for such factors are often arranged in ‘regulatory modules’, which contain several binding sites in close proximity, affecting transcription collectively. Such modules then act in a discrete manner, driving specific expression patterns as a single unit, largely independent of the nearby flanking DNA [144].

2.1.2 *Alternate mechanisms of transcriptional regulation*

In addition to functions carried out by transcription factors and their specific binding sequences, other mechanisms can also affect gene regulation. Structural characteristics of the DNA molecule [24, 103, 150] and accessibility of the DNA outside the histone complex [13, 36] have both been shown to have effects on gene expression patterns. For instance, left-handed Z-DNA structures occurring near the promoter have been known to regulate transcription [87, 103, 106]. Although Z-DNA conformations are physically unfavorable compared to the more common B-DNA conformations, Z-DNA structures can often be formed by certain nucleotide sequences, such as poly(GC), poly(CA), or poly(TG) dinucleotide repeats [103]. These repeat sequences and the resulting Z-DNA conformations are often found to inhibit transcription when occurring upstream of a given target gene [87, 106].

Accessibility of the DNA outside the nucleosome can also affect the transcriptional process [13, 36]. DNA tightly wound around the histone complex is often inaccessible to *trans*-regulatory elements, and therefore such regions are generally inactive [36]. Thus, displacement of the nucleosome from the DNA is essential to allow for the transcriptional process to occur. Various mechanisms allow for displacement of the histone complex, such as nucleosome remodeling, where the histone complex is altered via interacting proteins, allowing other transcription factors access to the DNA [66]. Other mechanisms, such as post-translational alterations of the histone molecules by phosphorylation, ubiquitination, or acetylation have also been known to

affect nucleosome packaging [117]. While the mechanisms mentioned above generally involve *trans*-regulatory factors, various sequence elements within the DNA themselves can inhibit nucleosome compaction. For instance, poly(T) thymine tracts have been shown to promote DNA bending, thereby displacing the nucleosome from the DNA molecule [28, 48, 111]. Although most studies regarding transcriptional regulation focus primarily on transcription factors and their binding sites, DNA accessibility plays a crucial role in gene regulation [36] and can also offer a means by which to detect active promoter regions within the genome [13].

2.1.3 Regulatory sequence analysis

With the recent availability of genomic DNA sequences, researchers have begun to use sequence analysis to gain an understanding of various biological processes. Of major interest is the ability to distinguish functional genomic elements from the non-functional nearby flanking sequences. In terms of transcriptional regulation, this involves the ability to discover sequence motifs to which transcription factors preferentially bind. The most common approach used to predict transcription factor binding sites involves searching for statistically overrepresented motifs upstream of the TSS [3, 17, 18, 32, 62, 67, 93, 105, 113, 121, 126, 127]. However, regulatory motif detection through computational methods poses a difficult problem. Regulatory motifs are usually short (e.g., 10 bp or less) and can be highly degenerate. In most cases, motif detection methods search for overrepresented motifs within a group of co-regulated genes, assuming that the same regulatory motif is present upstream of all, or the majority, of these genes [3, 32, 62, 67, 93, 105, 121]. However, it is generally not the case that co-regulation implies motif co-occurrence. For instance, a group of genes in a common pathway can show similar expression patterns, although genes upstream in the pathway can be regulated by different factors than those downstream in the pathway. In addition, statistical approaches used to search for

co-occurring motifs are diverse in methodology, and can be highly sensitive to user-defined parameters [71, 124]. As a result, different motif detection algorithms produce varying predictions, even when applied to the same data sets. In general, assessments of such motif detection programs have shown that this approach has very limited efficacy, whether applied to biological or to simulated data sets [124].

Other approaches have also been proposed to predict DNA sequence motifs that play a functional role in gene regulation. For instance, genome-wide scans for over-represented motifs within the regulatory region have also been proposed [17]. In such models, a ‘dictionary’ of common motifs is inferred from a large number of proximal promoter regions across the genome. Another common approach is to conduct cross-species comparisons in a method termed ‘phylogenetic footprinting’, which predicts regulatory motifs according to sequence conservation across species [60, 61, 68, 131]. However, it has been shown that many functional binding sites are not conserved across species [72, 91, 97]. This observation can often be explained simply by evolutionary changes within the binding site [97], or else by regulatory element turnover, where orthologous target genes share the same, yet non-homologous, regulatory element [72, 91]. Thus, considerable challenges remain regarding the detection of functional regulatory elements.

2.2 Available genomic sequence data

2.2.1 *RefSeq gene annotations*

Biological sequence analysis is facilitated by the ever-growing amount of available genomic sequence data. Studies regarding transcriptional regulation are often conducted using available TSS data annotations, which allow the user to analyze regulatory regions near the start of transcription. The RefSeq database in particular [76, 101] provides a large collection of high-quality gene annotations, including TSS annotations of mRNA transcripts. RefSeq annotations are maintained and made

freely available via Entrez [109] through the National Center for Biotechnology Information (NCBI) (<http://www.ncbi.nlm.nih.gov/>), and can also be accessed using the UCSC Genome Browser (<http://genome.ucsc.edu/>) [58].

From a computational point of view, obtaining reliable genomic information is crucial to conduct DNA sequence analyses. In particular, accurate TSS annotations are essential to study transcriptional regulation in cases for which the location of a given regulatory element within the promoter is of interest. With the advent of large-scale databases containing an albeit comprehensive list of TSS locations, it is now possible to conduct promoter sequence analyses on a genome-wide scale. In Chapter 3, we describe an approach that utilizes known TSS locations in order to study location-specific *cis*-regulatory elements within and near the promoter, effectively utilizing such high-quality TSS annotations.

2.2.2 Genome projects across diverse vertebrate species

There is now a growing amount of genome sequence data for a diverse number of organisms, ranging from human to bacteria. Even within vertebrates, complete genomes are available for more than two dozen organisms, including eighteen mammalian species [58]. Such a wide variety of both closely- and distantly-related lineages has allowed for evolutionary studies across a large array of organisms. In many cases, whole-genome alignments across multiple lineages are currently available [85], facilitating comparisons between species. Although the amount of coverage of the genome varies greatly across species, rising technology is rapidly increasing both the amount and quality of genomic sequence data.

Genome comparisons have provided a means by which to assess both similarities as well as differences between organisms. For instance, cross-species genome comparisons have estimated that approximately 80% of all human genes contain clearly identifiable orthologs in mouse [86], and over half contain orthologs in opossum [83].

Sequencing and analysis of the chimpanzee genome has shown that the human and chimp genomes are surprisingly similar, with each species accumulating substitutions at only 1% of all nucleotide sites since their divergence [120]. In particular, the vast majority of protein-coding genes shared within primates are strikingly similar, with $\sim 29\%$ of all genes differing only at 2 or less amino acid residues between human and chimp. This finding supports the claim made by King and Wilson [64] that evolutionary changes in phenotype can be largely explained through changes in gene expression patterns, rather than modifications within protein-coding regions.

These genome comparisons provide an initial survey of similarities as well as possible differences between species. Of even greater significance, however, is the availability of these genome sequences for further in-depth studies regarding the nature of molecular evolution. Prior to high-throughput sequencing technology, the study of molecular evolution was often limited to theoretical models, without the ability to conduct large-scale analyses due to the small amount of available data. The recent availability of such data thus allows for such models to be implemented on large high-quality data sets. Genomic sequence data therefore provide researchers with the ability and motivation to conduct cross-species analyses to study the process of molecular evolution.

2.3 Locational specificity in *cis*-regulatory elements

2.3.1 *Promoter architecture*

It has long been recognized that certain *cis*-regulatory elements function at specific locations within or near the promoter [15, 148]. Originally, only a few elements were known to exhibit locational specificity, such as the TATA-box (TBP binding element), the GC box (SP1 binding element), the CAAT-box (NFY binding element), and the Inr sequence [15]. The most well-studied of these elements are the TATA-box and the Inr sequence. Both are known to function at very precise locations within the core

promoter, defining the location at which transcription is initiated [79, 107, 108, 148]. The TATA-box plays a crucial role in transcription for many of its target genes, recruiting RNA polymerase II via protein-DNA interactions with TFIID, one of the GTFs within the basal transcriptional complex [142, 148]. The TATA-box is most often found approximately ~ 30 bp upstream of the start of transcription, and it is estimated that between 10-25% of all mammalian promoters contain a functional TATA-box element [107, 148].

Although the TATA-box is commonly found within many promoters, it is conspicuously absent from others. This is particularly true for housekeeping genes [38, 66, 148], i.e., those that display consistent expression patterns across tissues and external environments. In many genes that lack a TATA-box, the start of transcription can be accurately placed via the Inr sequence, which occurs directly across the start of transcription [148]. The common Inr sequence, present in approximately $\sim 46\%$ of all eukaryotic genes, generally controls the location of transcriptional initiation by binding two transcription factors (TAF1 and TAF2) within the TFIID protein complex [19, 148]. Interestingly, TBP, the factor within the TFIID complex that generally binds to the TATA-box, is also recruited to the basal transcriptional complex in most TATA-less promoters. However, in the absence of the TATA-box, TBP itself does not bind to the DNA, but is instead recruited to the basal transcriptional complex through protein-protein interactions with other GTFs [19, 79, 148].

The locations of the TSS within promoters lacking both a TATA-box and the Inr element appear to be largely determined by other protein factors outside the basal transcriptional complex [66, 137]. These factors, such as the GC box and the CAAT box, then interact with the basal transcriptional complex and position this complex in general approximation to the start of transcription [66, 148]. TATA-less promoters are often located within CpG islands, and thus the dinucleotide composition of these promoters is qualitatively different than those that possess a functional TATA-box

[148]. The mechanisms by which TATA- and/or Inr-containing promoters operate appear to be distinct from TATA- and Inr-less promoters [55, 56, 107]. For instance, the location of the TSS within promoters containing a TATA-box or Inr element is generally well-defined, and is accurately located at a specific nucleotide site [56]. In contrast, promoters lacking both elements are most often found within housekeeping genes and can contain several initiation sites, interspersed across potentially 50-100 nucleotide sites [55, 107]. Thus, certain authors have emphasized the diversity of promoter architecture across genes, which can be placed into one of several categories according to the makeup of their promoters and the mechanisms by which they operate [55, 56, 115, 148].

2.3.2 Locational-specificity within regulatory motifs

Locational specificity had originally been documented for only a small handful of regulatory motifs [15], and only recently has the prevalence of location-specific overrepresentation been studied among other *cis*-regulatory elements. Different assessments of the prevalence of locational specificity among regulatory elements have produced conflicting results. The predicted number of motifs exhibiting locational specificity ranges from only nine non-redundant 8mer predictions made by FitzGerald et al [38] to 1,226 unclustered 8mer predictions made by Tharakaraman et al [119]. Although, in many cases, there exist similarities in the approaches taken by various studies, subtle differences in methodology can often lead to varying results.

There are two major challenges to overcome when predicting locational-specificity among putative regulatory elements. The first obstacle is the difference in the size of the region at which a motif is overrepresented. Some regulatory motifs are constrained to very precise locations within the promoter, while others can be found overrepresented across large areas of the regulatory region. For instance, the Inr sequence functions at a single nucleotide site overlapping the TSS [15, 56, 148], and

the TATA-box is found overrepresented at only a few nucleotide sites upstream of the TSS [56, 79]. In contrast, most location-specific regulatory elements, such as the GC box and the CAAT-box, are found overrepresented across much broader regions of the proximal promoter. Such motifs can often be overrepresented across ~ 50 -150 nucleotide sites, although the number of sites can vary greatly between regulatory motifs [15, 149]. With only one exception [119], most studies do not take such differences into account. Instead, most studies generally search for location-specific overrepresentation within several discrete windows of predefined width (e.g., ~ 20 -25 bp) [38, 128, 145, 146]. Studies that take this low-resolution approach have limited sensitivity to many instances of location-specific overrepresentation, particularly for cases in which a motif is found constrained to a highly specific location, or else overrepresented across a broad range of the regulatory region. For instance, none of the studies using this method could predict locational specificity for the *Inr* sequence *de novo* (Section 3.4.1) [38, 128], despite the fact that the *Inr* sequence has been known to exhibit locational specificity for more than two decades [15]. Similar difficulties were encountered for the TATA-box. For example, Vardhanabhuti et al [128] were not able to predict location-specific overrepresentation of this well-known element at the location at which it is known to function.

A second difficulty encountered during location-specific motif prediction is the effect of dinucleotide fluctuations close to the TSS. Many promoters occur within CpG islands [56, 148], and thus GC dinucleotide content rises significantly near the TSS [11, 38, 55]. Studies that do not account for locational fluctuations in dinucleotide content therefore generally predict a disproportionate percentage of GC-rich motif predictions [119]. We show in Section 3.4.2 that the majority of these predictions are simply due to the rise of GC content near the TSS, and they generally do not comprise true *cis*-regulatory elements.

2.3.3 Assessments of locational specificity

There are three major facets regarding location-specific motif predictions. The first is to determine which previously known TFBSs exhibit locational specificity. This allows us to predict which transcription factors are likely to be involved in positioning the basal transcriptional complex within the promoter. We show in Chapter 3 that many previously known regulatory elements exhibit locational specificity, although their locational overrepresentation has not previously been documented. Second, since certain regulatory proteins are known to function at particular locations within the proximal promoter, a given model can be validated according to the presence of the protein binding motifs within its list of predictions. As mentioned in the previous section, many approaches designed to predict locational specificity cannot detect even well-known location-specific motifs, and therefore fail this test in certain cases [38, 128, 146]. A third purpose for analyzing locational specificity is to predict previously unknown sequence elements that may play novel roles in transcriptional regulation. Predicting novel regulatory motifs *de novo* therefore furthers our understanding of transcriptional regulation and the sequence elements controlling gene expression.

One common approach taken by studies regarding locational specificity is to scan for location-specific overrepresentation within either previously known regulatory elements or across motifs predicted using alternative methods [81, 128, 145, 146]. The latter case usually includes motifs predicted using either standard motif detection methods [145] or cross-species conservation [146]. These approaches have certain drawbacks. For instance, although the prevalence of locational-specificity within a set of known regulatory elements can be assessed using these approaches, it is limited by our current knowledge of TFBSs or the efficacy of orthodox motif detection methods. By the nature of this approach, novel motifs cannot be predicted using locational overrepresentation as a criteria for functionality.

The second approach is to predict positionally overrepresented motifs *de novo*, scanning across a comprehensive list of k-mers and inferring functionality according to locational-specificity itself [38, 128, 119]. In this second approach, no prior knowledge is required concerning the sequences of potential location-specific regulatory motifs. The work described here uses this approach (Chapter 3), and provides comparisons to the results of various other studies using similar, but distinct, methodologies.

2.4 Mutual relationships between regulatory elements

2.4.1 Combinatorial regulatory element relationships

As discussed previously in Section 2.1.1, transcription is often controlled by multiple transcription factor proteins acting collectively within multi-protein complexes [4, 10, 33, 66, 144]. Such complexes are generally bound to the DNA within regions called ‘*cis*-regulatory modules’ (CRMs), each of which contains several binding sites that recruit one of the many protein components [66, 144].

Due to the combinatorial effects of transcription factor proteins upon their target genes, several studies have analyzed regulatory element combinations within predicted CRMs [2, 88, 110, 114, 151]. Such studies have analyzed densities of binding sites within small sequence windows across the regulatory region (e.g., 200 bp [2]). There are various applications of CRM detection, such as predicting differential gene expression patterns [5], analyzing binding site combinations within co-expressed genes [110], *de novo* prediction of individual regulatory elements and their functional relationships [88, 151], or simply the identification of CRM locations within the genome [2, 114]. Approaches used to detect CRMs can vary widely between studies. Some methods have been applied to a small set of co-regulated genes without prior knowledge regarding the binding site sequences [88, 110, 151], while others use previously known binding motifs, either within co-regulated genes

or across the entire genome [2, 114].

2.4.2 Distance relationships between regulatory motifs

In most studies, functional regulatory motif relationships are studied only according to binding site densities within discrete windows representing a CRM. However, this represents an over-simplification of protein-protein interactions. Such interactions are not only dependent upon proteins binding within the same vicinity of the DNA molecule, but are also driven by more complex structural conformations of both the proteins as well as their orientation with respect to the DNA [63, 66, 144]. We might therefore imagine that certain distances between the binding sites are more favorable than others to promote protein interactions.

A few previous studies have incorporated models that analyze distance preferences between specific binding motifs, rather than an all-inclusive density of binding site occurrences within a single sequence window [5, 16, 88, 114, 128]. In the simplest case, spatial preferences are measured using maximum-distance models, where regulatory motif-pairs are assumed to function collectively when they occur within some maximum distance from each other [5]. Although this approach is easy to implement, it has had little success and does not significantly improve efforts to study gene expression patterns [5]. A similar approach was used by Sinha et al [114], who assumed a co-occurrence probability distribution that decreased exponentially according to the distance between two motifs. However, this model does not accurately reflect true binding site relationships, since binding sites in very close proximity are unlikely to promote protein-protein interactions. This is due to the fact that protein binding occurs across a larger portion of the DNA than that of the canonical sequence specific binding motif [66], and thus the inherent size of the proteins require some minimal buffer between motifs to allow for protein interactions to occur.

A more realistic approach is to use a ‘sliding window’ method, which measures inter-motif distance preferences according to motif co-occurrences at different distances within a window of pre-defined width [16, 128]. This model can be applied using previously known transcription factor binding motifs or those predicted using alternative methods [16], or it can also be used to predict *cis*-regulatory elements *de novo* by comprehensively searching across all pairs of k-mer motifs [128]. In this latter case, no prior knowledge is necessary regarding the sequences of the regulatory motifs involved. Thus, this approach can be used to predict functional motif-pair relationships between both known as well as previously undocumented motifs.

2.4.3 Periodic distributions of inter-motif distance preferences

All previous approaches designed to study spatial preferences between sequence motifs assume unimodal distributions between regulatory elements [5, 16, 88, 114, 128], whether using a maximum-distance approach or the sliding window approach. However, some experimental evidence has suggested that inter-motif distance preferences may in fact be periodic [53, 69, 135, 144]. For instance, binding protein-pairs often prefer to occur in a particular orientation to each other relative to the turn of the double helix [69, 135, 144]. Such requirements thus produce phased distance preferences corresponding to the turn of the DNA molecule. Other periodic distributions have been associated with wrapping of the DNA around the histone complex [53]. Thus, we would expect to see more than one preferred distance between motifs, each separated by non-preferred distance intervals. Therefore, unimodal approaches are likely to be oversimplified, as they do not account for periodic distance preferences between motifs. Although multi-modal distribution approaches are less straightforward to implement, we show in Chapter 4 that such models fit the data more accurately, and that it is a necessary extension of previous approaches designed to detect spatial preferences between motifs.

2.5 Evolution of transcriptional regulation

2.5.1 Significance of gene expression on the evolving organism

Gene regulation acts as a major contributor to phenotype, and therefore evolutionary changes in gene expression can have significant impacts upon a given lineage [1, 20, 66, 116, 144, 143]. There is considerable controversy over the relative importance of *cis*- and *trans*-element modifications upon species morphology [21, 49, 140, 143]. Presently, more is known about protein-coding mutations [140], although a growing body of evidence has shown that changes in *cis*-regulatory elements act as major contributors to variation across species [21, 143]. One recent survey of protein binding events for three liver-specific transcription factors (HNF1A, HNF4A, HNF6) show that such binding events tend to vary across human and mouse, and that these differences are caused primarily by evolutionary changes in *cis*-regulatory sequence [139]. In contrast, *trans*-effects in binding proteins, histone remodeling, and cellular environment play only secondary roles. *cis*-regulatory changes between human and stickleback have also been associated with Kit ligand expression, consequently affecting pigmentation [84], and species-specific differences in malaria resistance among primates have likewise been attributed to changes in *cis*-regulatory regions [125].

The fact that many differences in gene expression arise from *cis*-regulatory effects suggests that differences in expression across species can be studied at the DNA sequence level. Still, as outlined in the following sections, certain complexities pose a major challenge to applying sequence analysis approaches towards such problems. In Chapter 5, we outline an approach that uses cross-species comparisons to study *cis*-regulatory elements and evolutionary changes in their consensus sequences. This approach effectively circumvents such complexities, and further explains the biological mechanisms that have previously presented challenges to the study of regulatory

element evolution.

2.5.2 *Preservation of gene expression patterns versus sequence conservation*

Recently, there has been a growing body of evidence that contradicts previous assumptions that sequence divergence is correlated with changes in gene expression [25, 37, 140]. Surprisingly, it has been shown that gene expression patterns are highly conserved even across very distantly related species, such as between human and *fugu* or zebrafish [25, 37]. Interestingly, many of the genes whose expression patterns have been conserved are not ubiquitously expressed housekeeping genes. Instead, the majority of genes with conserved expression tend to be tissue-specific, performing unique functional roles within the cell [25]. Yet, despite the high frequency of conservation in the expression patterns of many genes, conserved expression patterns correlate poorly with the amount of sequence conservation [25, 37, 44]. In some cases, changes in genomic sequences are caused by regulatory element turnover, where the same regulatory element targets orthologous genes, although the regulatory element sites themselves are not homologous across species [30, 37, 44, 91]. For instance, studies conducted across various *Drosophila* species have shown that individual regulatory element rearrangements are extremely common within developmental *cis*-regulatory modules [44]. This observation suggests that ‘phylogenetic footprinting’ methods that scan for conserved regulatory elements are not well-suited to predict conservation of expression patterns across species.

It has further been shown that, in many cases, conserved expression patterns can exist in the absence of transcription factor binding events [12, 91, 140]. For instance, many, and in some cases the majority, of transcription binding events in liver-specific genes do not overlap between human and mouse [91]. Comparative analyses within different yeast species have also shown that most transcription factor binding events are species-specific, even in cases for which gene expression patterns

across orthologous genes have been preserved [12]. The apparent separation between gene expression patterns and sequence divergence has presented a major obstacle for sequence analysis approaches regarding the evolution of gene regulation. To date, such studies have met with little success, with notably high false-positive and false-negative rates [31].

2.5.3 Evolution of *cis*-regulatory element consensus sequences

It has been commonly assumed across a broad range of studies that the preferred binding sequences of regulatory proteins remain largely unchanged over time [31, 44, 60, 61, 72, 128, 141, 146]. For instance, studies using regulatory element conservation to predict functional sequence elements naturally assume that preferred regulatory motifs are identical across species [60, 61, 141, 146]. Approaches focusing upon regulatory element turnover likewise make this assumption [44, 72, 91], and in some cases these studies focus specifically upon regulatory motifs whose binding proteins are functionally and structurally conserved [91]. Predicted losses and gains of *cis*-regulatory elements are also made under the assumption that regulatory consensus motifs are identical across species [31], disregarding the presence of TFBSs whose binding sequences differ between lineages.

Some studies have explicitly argued that protein binding affinities remain unchanged over time [9, 25], with Berger et al [9] asserting that homeodomain binding affinities have remained static since the divergence of all animals, including mammals, *Drosophila*, and *C. elegans*. This claim rests solely upon the fact that binding modules within homeobox factor proteins have been largely conserved between lineages, and that computational predictions suggest similar binding sequences across a wide range of animals. However, their computational predictions were only experimentally verified for one homeodomain in *Drosophila* and *C. elegans*, out of a total of 168 homeodomains existing in mouse [9]. Their approach, by design, predicts binding

motifs according to sequence similarities at 15 specific amino acid sites within the homeodomain (~ 60 residues total). In particular, homeodomains matching at all 15 sites across species automatically result in identical predictions of binding affinity. As the authors note, there is a higher amount of variation in the 168 homeodomains present within the mouse genome than there is variation across species, and thus many homeodomains between mouse and *Drosophila* match exactly at all 15 sites. It is, however, unlikely that binding affinities can be characterized completely using only 15 specific amino acid residues, particularly when disregarding the remaining ~ 45 amino acids within the homeodomain. In addition, homeobox proteins are known to be susceptible to changes in structure according to external stimuli, with certain homeobox factors altering their structure and binding preferences according to the presence of interacting protein factors [78, 138].

Despite the claims by Berger et al that binding affinities of nearly all homeodomains have remained constant since the divergence of all animals, no previous study has provided strong evidence supporting the hypothesis that regulatory elements have remained constant even within closely-related lineages. Our observations discussed in Chapter 5 provide evidence to the contrary, showing that a surprisingly large fraction of regulatory motifs have been subject to significant modifications even since the divergence of closely-related species.

The dichotomy between conservation in sequence and the conservation of expression patterns has previously presented a significant challenge to studies regarding the evolution of gene expression. Our approach directly addresses this separation between expression and sequence conservation under the hypothesis that changes in regulatory element sequences can be systematically modified across the genome. Modifications in the preferred binding motifs of orthologous transcription factors may then produce functional conservation in expression, although the *cis*-regulatory element sequences themselves may be altered. Our results may then explain one

of the mechanisms by which modifications in sequence can still preserve gene expression, providing one interpretation of the paradoxical observations regarding the nature of regulatory element evolution.

Motif Locational Functions

Many *cis*-regulatory elements play a crucial role in determining the location at which transcription is initiated [79, 66]. Several known regulatory motifs have previously been shown to exhibit locational specificity within the proximal promoter, such as the Inr sequence and the TBP, SP1, and NFY binding sites [38, 66, 128, 145, 146]. However, the extent to which location-specific overrepresentation occurs among regulatory motifs is not well understood. Previous approaches designed to detect location-specific overrepresentation [38, 119, 145, 146] generally neglect the fact that dinucleotide content fluctuates near the start of transcription, and thus such methods are typically sensitive to rises in GC content near the TSS. In addition, previous studies use low-resolution approaches [38, 128, 145, 146], decreasing sensitivity to instances of overrepresentation occurring at very precise locations. Such approaches often fail to detect well-known motifs exhibiting locational specificity, such as the TATA-box [128] and the Inr sequence [38, 128].

In this chapter, we assess the prevalence of location-specific overrepresentation among *cis*-regulatory elements and utilize this characteristic to predict novel *cis*-regulatory element candidates. We offer a means by which to study locational specificity that accounts for dinucleotide fluctuations within the promoter and also

considers the data at single-site resolution, increasing sensitivity to locational overrepresentation occurring both across broad ranges of the proximal promoter as well as those occurring at only one or a few sites.

We provide a general model, denoted as the ‘motif positional function’ (MPF) model, that can be used to detect spatial preferences relative to any given reference point. In this work, we describe two such models: the ‘motif locational function’ (MLF) model, as well as the ‘motif relational function’ (MRF) model. The first model determines locational overrepresentation within the proximal promoter relative to the TSS. We also discuss the MRF model, which is designed to detect spatial preferences between pairs of motifs. In this chapter, we focus on the former model; methods and analyses pertaining to the second (MRF) model are discussed in the following chapter.

3.1 Motif locational functions (MLFs) predict locational specificity in *cis*-regulatory motifs

3.1.1 Overview of the MLF model

In contrast to previous studies, we determine location-specific overrepresentation at single-site resolution using regression analysis, allowing the data to be considered collectively across each position. MLFs are modeled using a continuous function $g(x)$, whose values represent the underlying probability of occurrence according to position x . Here, x denotes the (single) nucleotide position relative to the TSS. Thus, $x = -20$ refers to the site 20 bp prior to the TSS, and $g(x) = g(-20)$ represents the probability density of motif occurrence at this site.

In our model, $g(x)$ is given as the sum of the background frequency $C(x)$ and a contribution of location-specific overrepresentation $H(x)$. Thus, $g(x) = C(x) + H(x)$ (Appendix A.1). Locational overrepresentation, represented by $H(x)$, is modeled using an unnormalized Gaussian term, incorporating a ‘peak’ into the MLF:

$$H(x) = a \cdot \exp \left[-\frac{(x - \mu)^2}{2\sigma} \right] \tag{3.1}$$

where a , μ , and σ are free parameters, and vary between motifs.

The mean (μ) of the Gaussian term represents the central location of overrepresentation, i.e., the x -value at which the peak of overrepresentation reaches its maximum value. This maximum value is represented by the coefficient a , which gives the height of the peak over the background frequency of occurrence (note that $H(\mu) = a$). The standard deviation (σ) then reflects the ‘width’ of the region at which the motif is overrepresented. Figure 3.1a illustrates this scheme for two motifs exhibiting locational overrepresentation, namely, the TBP binding site (TATA box) and the SP1 binding site (GC box). These motifs are found overrepresented 30 and 65 bp prior to the TSS, respectively ($\mu = -29.6$ and $\mu = -65.3$). The TATA-box is found overrepresented at only a few sites within the promoter, thus producing a small σ -value ($\sigma = 1.9$), while the GC box is overrepresented across a much broader range upstream of the promoter ($\sigma = 52.2$).

For a given motif, we predict spatial preferences using a likelihood ratio test (F -test) (Appendix A.4). Namely, we compare the model assuming no locational specificity to that allowing for location-specific overrepresentation. The former is modeled by setting $H(x)$ to zero, and thus the underlying frequency of occurrence $g(x)$ simply equals the background frequency $C(x)$. We compare this model to the one where $H(x)$ takes on non-zero values, allowing for location-specific overrepresentation. Comparing the log-likelihoods given each of these two models produces a p -value reflecting the significance of locational overrepresentation. Locational specificity is then predicted for motifs for which the p -value falls below a given threshold.

3.1.2 Accounting for dinucleotide fluctuations within the promoter

Our model accounts for fluctuations in dinucleotide frequencies across the promoter by allowing the values of the background frequency $C(x)$ to vary according to location (Appendix A.2). For instance, GC content rises near the start of transcription [38], and therefore the background frequency $C(x)$ of GC-rich motifs likewise increases close to the TSS. In order to determine the background frequency, we first

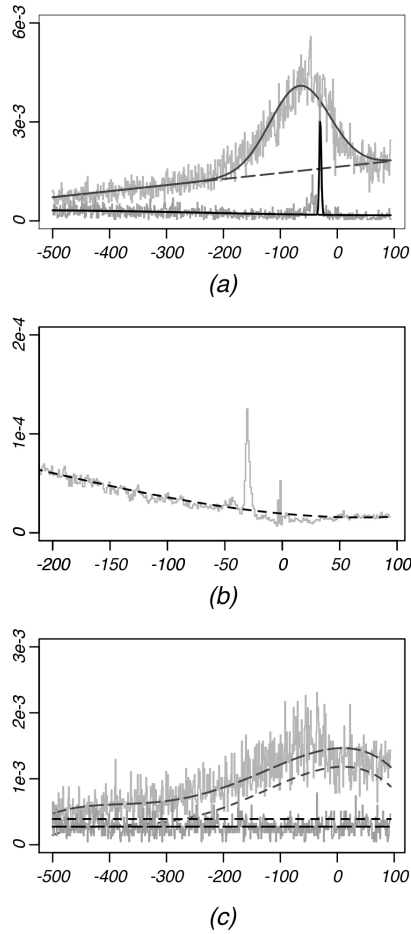


FIGURE 3.1: Raw data and MLFs of four example motifs. x -axis values denote the position within the promoter, where $x = 0$ defines the location of the TSS. The y -axis represents the frequency of occurrence. Solid plots represent the resulting MLFs ($g(x)$), long dashes show the background frequency of occurrence $C(x)$, and short dashes show the expected frequencies derived from dinucleotide composition ($c(x)$). (a) MLFs of the SP1 (gray) and TBP (black) binding sites; significant amounts of location-specific over-representation are predicted for both. (b) Expected frequencies of the TBP-binding site. Each data point is derived according to the dinucleotide composition at each position. (c) Two motifs (GGGCGC, gray; TGCTTC, black) without locational specificity. Note that without positional enrichment ($H(x) = 0$), the MLF $g(x)$ is the same as the background frequency $C(x)$. A comparison between $C(x)$ and $c(x)$ illustrates the ability of the background model to account for uniformly distributed over- and under-representation with respect to the expected frequency (according to dinucleotide composition). Fluctuation within the gray plot is attributed to the dinucleotide makeup of the promoter (note $c(x)$); this motif is therefore not predicted to exhibit locational specificity.

estimate each motif’s expected frequency across the sequences. This expected frequency is determined strictly according to the dinucleotide makeup of the motif as well as the dinucleotide makeup of the promoters; this frequency is denoted as $c(x)$. The values of $c(x)$ are allowed to vary by position according to the dinucleotide makeup of the regulatory region, thus accounting for location-specific changes in dinucleotide composition. However, we distinguish between the ‘background’ and ‘expected’ frequency of occurrence (given by $C(x)$ and $c(x)$, respectively), as many motifs are either over- or under-represented with respect to their dinucleotide composition. Thus, the background frequency $C(x)$ allows for uniformly distributed over- and under-representation. Both $c(x)$ and $C(x)$ are important components of our model. Namely, we must allow for differences between the expected and observed frequency of occurrence while still incorporating dinucleotide fluctuations into the background frequency of occurrence. Thus, although the background frequency is allowed to deviate from the expected frequency, $C(x)$ is restrained to mimic the ‘shape’ of $c(x)$. This preserves the expected fluctuations according to dinucleotide composition. For instance, for motifs whose frequencies are expected to vary according to location, rises and drops in $C(x)$ are restricted to conform to those of $c(x)$. In contrast, motifs expected to occur at a constant frequency across the region (i.e., the values of $c(x)$ are uniform across all positions) likewise have a constant value for $C(x)$.

We model $c(x)$ in a continuous fashion using a polynomial function (i.e., $c(x) = \sum_k b_k x^k$). This function is determined by conducting linear regression on the set of data points representing the expected frequency of occurrence at each site. Fitting the function $c(x)$ to these data points then gives the expected frequency of occurrence. An example is illustrated in Figure 3.1b, which shows the raw data and resulting function $c(x)$ for the TATA-box. We note that sharp rises in the observed dinucleotide frequencies at a particular location are not directly incorporated into $c(x)$, but instead remain outlier points after fitting this function to the data. This is an important aspect of our model, as overrepresentation of a motif can itself cause

rises in dinucleotide frequency [128]. Incorporation of such rises into $c(x)$ would therefore obscure the distinction between the signal and the background frequency at this location.

The ability of the background model to incorporate uniformly distributed over- and under-representation is illustrated in Figure 3.1c. Here, we show the MLFs for two motifs that do not exhibit locational specificity. We note that although the gray plot fluctuates according to location, this would be expected according to the dinucleotide frequencies within the promoter (note the fluctuations in $c(x)$). Thus, our model does not predict this motif to exhibit biologically relevant locational specificity, as rises in occurrence frequency near the TSS are simply a byproduct of its high GC content.

3.2 The MLF method predicts location-specific overrepresentation for many motifs within human promoters

In order to determine which motifs exhibit locational overrepresentation within human promoters, we scanned for locational specificity across all 6-mer motifs on a set of non-redundant RefSeq human promoters [76, 101] collected from the UCSC Genome Browser (<http://genome.ucsc.edu>) [58]. The data set consisted of 20,609 sequences, each comprising the region 500 bp upstream and 100 bp downstream of a known TSS. As expected, the vast majority of 6-mer motifs did not exhibit locational specificity within the promoter data set. However, a few motifs showed highly significant locational overrepresentation, with 106 6-mers exhibiting locational specificity at a significance level of $p < 1e-25$.

3.2.1 *Predicting locational overrepresentation according to results from control data sets*

To compare these results to those of a control data set, we repeated the analysis on a set of intergenic sequences, each comprising the 600 bp interval starting 2 kb upstream of a known TSS. Very few motifs were predicted to exhibit positional enrichment in this control data set, with less than 1% producing p -values under $1e-5$.

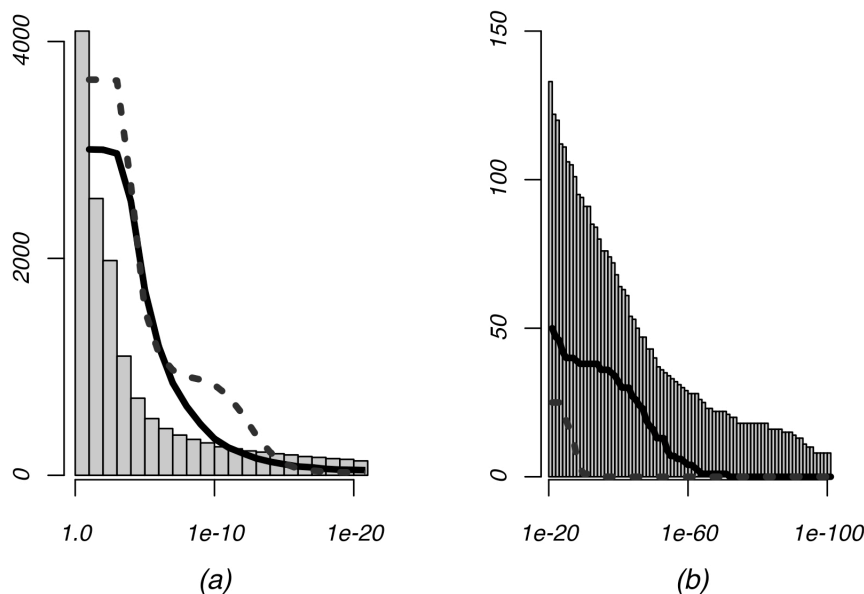


FIGURE 3.2: Results from the comprehensive MLF analysis. Histograms show the cumulative number of 6-mers with location-specific overrepresentation according to their p -value significance in human promoters. The plots give the number of 6-mers producing p_{sim} -values under the given thresholds during simulation analyses, where p_{sim} represents the most significant p -value for each individual 6-mer across 100 simulated data sets. Solid plots refer to simulations conducted according to the dinucleotide frequencies across each position within the promoter, while the dotted lines represent those generated using mono-nucleotide frequencies. p -value thresholds above $1e-20$ are shown in (a), while the contrast between the results of the human versus simulated analyses for $p < 1e-20$ is illustrated in (b). Note that the dinucleotide-generated simulated data sets produced a significantly larger number of predictions than the mono-nucleotide-generated simulated data, while the real human promoters produced more predictions than either of the control data sets.

We then tested our model on two types of simulated data sets. The first was generated by considering the observed mono-nucleotide frequencies at each site, while the second was produced using dinucleotide frequencies at each position. One hundred data sets of both the mono- and di-nucleotide simulations were generated, with each individual data set comprising the same number of sequences as the human promoter data. For each type of simulation, we scanned for locational overrepresentation across all 100 data sets, recording the most significant p -value for each individual 6-mer. This p -value, denoted as p_{sim} , was thus unique to each 6-mer. Re-

sults of these analyses, as well as those for the real data analyses, are shown in Figure 3.2. Significantly more predictions were made during the human promoter analysis than either of the simulation analyses. We also note that the dinucleotide-generated data sets produced more predictions than those produced using only mono-nucleotide frequencies, suggesting that our use of a dinucleotide-based background model leads to a more conservative significance criteria.

We used the results of both the intergenic sequence analysis as well as those of the simulated data sets to set the prediction criteria for locational specificity within the human promoters. The lowest p -value produced from the intergenic sequence analysis was found slightly under $1e-15$; motifs above this threshold were excluded from the list of predictions. The remaining motifs producing p -values under their p_{sim} -values times a stringent multiple hypothesis correction factor of $1e-5$ were then predicted to exhibit locational preferences in the human RefSeq promoters. Thus, the prediction criteria ($p < p_{sim} \times 1e-5$) was unique to each 6-mer, subjecting motifs with lower p -values within the simulated data sets to a more stringent threshold.

The final list of predictions contained 166 6-mer motifs, representing 4% of the total number of possible 6-mers. Despite our stringent prediction criteria, the majority of the top-ranked motifs from the RefSeq promoter analysis were not filtered due to the simulation analyses. Out of the 50 top-ranked motifs exhibiting locational overrepresentation within the real promoters, only one motif did not pass the p -value threshold determined as above.

3.2.2 Motif clustering predicts locational overrepresentation for both known and putatively novel cis-regulatory elements

We used the list of location-specific 6-mers to generate consensus motifs with degenerate sites and flexible lengths. Motifs were clustered computationally according to sequence similarity as well as the location and width of their locational overrepresentation (Appendix A.5). We then condensed each cluster into a single consensus sequence, generating a total of 48 consensus motifs exhibiting location-specific overrepresentation within human promoters. In order to test whether the predicted mo-

tifs overlapped with known regulatory elements, we compared our results to known transcription factor binding sites (TFBSs) in the TRANSFAC database [80] using STAMP [77]. Thirty-four of the motif clusters matched known *cis*-regulatory elements, comprising a total of twenty known binding sites within TRANSFAC as well as the Inr sequence element (Table 3.1). Several of the motifs predicted were previously known to exhibit locational specificity, including the TBP, SP1, NFY, CREB, ETS, NRF1, and MYC factor binding sites [38, 128, 145, 146]. We also predicted several additional motifs whose locational specificity had not been previously documented, including fourteen novel regulatory motif candidates, denoted as d1-d14. The location of overrepresentation for each of the predicted motif clusters is illustrated in Figure 3.3. Most of these motifs were found to be overrepresented close to the TSS, although a few were found farther upstream of the promoter. Motifs overrepresented far from the promoter were frequently found to be overrepresented over a large range of the regulatory region, as shown in Table 3.1. This is to be expected, as it is unlikely that a regulatory element enriched far from the promoter would be constrained to a highly specific location. We note the precision of the method to predict related clusters at the same location, such as the TBP binding site as well as the Inr sequence clusters.

Figure 3.4 shows the MLFs for six motifs with locational overrepresentation within the proximal promoter. The MLFs of the GC-rich NHLH1 and the ZFP161 binding sites are shown at the top of the figure. We note that the rise of GC content centers directly across the TSS, as indicated by the simulated data plots. However, the positional enrichment for each motif is found at other locations ($\mu = +33$ and -51 , respectively), indicating that the locational bias of these motifs is not due to dinucleotide fluctuations near the TSS. The putatively novel d3 motif comprises a homopolymeric thymine tract. Such poly(T) sequences are known to alter DNA conformation, thereby affecting transcriptional regulation by displacing the nucleosome from the DNA molecule [28, 48, 111]. Similarly, the novel d10 motif, comprising a CA-dinucleotide repeat, promotes left-handed Z-DNA conformations [87, 103, 106].

Location-specific motif clusters													
Human RefSeq data						Mouse RefSeq data			Previous studies				
Rank	p	TF	Consensus	μ	(σ)	Consensus	μ	(σ)	Fitz	Xi	Xie	Vard	
1	7e-179	SP1 (+)	AGGGGGCGGG	-68.3	(52.2)	GRGGGGGCGKG	-69.6	(42.8)	*	*	*	*	
2	2e-106	NFY (-)	CTSATTGGCT	-78.8	(42.7)	ATTGGC	-100.0	(16.1)	*	*	*	*	
3	1e-102	CREB	CGTGACGTC	-49.1	(39.0)	GTGACG	-44.5	(34.6)	*	*	*	*	
4	3e-102	ZEB1 (-)	CAGGTAAG	72.5	(31.6)	GGTAAG	71.6	(33.9)	*	---	---	*	
5	5e-96	YY1	GATGGCGG	31.9	(22.1)	TGGCGG	23.8	(16.7)	*	*	---	*	
6	6e-94	NFY (+)	AGCCAATCAG	-76.7	(40.7)	GCCAAT	-91.0	(21.9)	*	*	*	*	
7	5e-91	d1	GTGAGTG	69.2	(36.4)	GTGAGTG	70.1	(32.6)	*	---	---	*	
8	3e-90	NHLH1	CAGCGGCKGC	33.0	(40.9)	RGCGCGC	32.6	(44.4)	---	---	---	*	
9	3e-87	SP1 (-)	CGCCCC	35.0	(32.2)	GCCCC	-66.3	(33.7)	*	*	---	*	
10	2e-83	ETS (+)	ACCGGAAGTG	-25.9	(32.3)	GCCGGAAGTG	-33.5	(37.0)	*	*	*	*	
11	9e-83	TBP	ATATAAAR	-30.6	(1.9)	ATATAAARGC	-30.9	(1.7)	*	*	*	---	
12	4e-74	SP1 (-)	GCCCCKCCCC	-76.2	(45.0)	SCYCKCCCC	-78.7	(51.7)	*	*	*	*	
13	7e-65	REST	CRCCATGGA	52.8	(38.0)	CGCCATGGCY	50.4	(34.9)	*	---	---	*	
14	2e-58	ETS (-)	CACTCCGGT	-24.3	(32.2)	CTTCCGG	-16.5	(16.0)	*	*	*	*	
15	1e-54	HBP1	RCGTAC	-47.0	(37.4)	CGTCAC	-53.2	(39.5)	---	---	*	*	
16	3e-53	ZFP161	CGCGGC	-51.8	(95.0)	CGCGCGC	-32.6	(97.1)	---	---	---	---	
17	1e-50	d2	TCTGCTGCT	51.0	(33.5)	CTGCTGCT	53.1	(37.0)	*	---	---	---	
18	2e-48	YY1	CAAGATGG	22.9	(17.1)	CAAGATGG	14.5	(10.7)	*	*	---	---	
19	3e-46	d3	TTTTTT	-12.7	(11.3)	---	---	---	---	---	---	---	
20	3e-45	TBP	TWTATA	-29.9	(2.0)	ATATAW	-27.9	(1.8)	*	*	*	---	
21	5e-44	NRF1	RTGCGCA	-53.7	(59.8)	TGCGCA	-57.8	(46.8)	---	*	---	*	
22	5e-40	NRF1	GCGCATGC	-46.9	(38.0)	---	---	---	---	*	*	*	
23	9e-39	Inr	GCTCAGTCC	-4.0	(0.2)	TCAGTC	-2.2	(0.5)	---	---	---	---	
24	5e-37	MYC	CACGTG	-51.0	(50.7)	CACGTG	-53.3	(46.1)	*	---	*	*	
25	7e-35	ZIC2	CCCACCC	-131.0	(70.2)	†CCCCC	-117.6	(99.9)	---	---	---	---	
26	1e-32	d4	TCCTCCT	-71.4	(82.9)	†CCCTCC	-61.9	(32.8)	---	---	---	---	
27	8e-32	d5	GTGTGT	-325.6	(234.4)	TGTGTGT	-435.8	(212.5)	---	---	---	---	
28	1e-25	TBP	AAAAGG	-27.3	(1.3)	---	---	---	*	---	---	---	
29	2e-25	SRF	ATGGCC	53.6	(33.9)	GATGGC	26.9	(20.1)	---	---	---	---	
30	5e-23	SOX9	CAATGG	-80.1	(23.9)	WCCAATGR	-85.7	(40.1)	---	---	---	---	
31	2e-21	d6	GGCGTG	-62.5	(34.1)	---	---	---	---	---	*	---	
32	2e-21	GTF2IRD1	CTCCCTC	-111.0	(100.6)	†CCCTCC	-61.9	(32.8)	---	---	---	---	
33	3e-21	d7	AAAAAA	-165.0	(10.2)	---	---	---	---	---	---	---	
34	2e-20	MEF2	AAAAAT	77.3	(23.1)	AAAAATA	202.3	(78.0)	---	---	---	---	
35	4e-20	d8	GCGCTC	-120.6	(174.9)	---	---	---	---	---	---	---	
36	7e-20	d9	GCAGCA	47.5	(36.0)	GCAGCA	28.6	(15.4)	*	---	---	---	
37	1e-18	Inr	CAGTTG	-1.2	(0.5)	†TCAGTC	-2.2	(0.5)	---	---	---	---	
38	2e-18	Inr	GTCACT	-3.0	(0.1)	---	---	---	---	---	---	---	
39	2e-18	d10	ACACACA	-12.6	(23.7)	---	---	---	---	---	---	---	
40	3e-18	TBP	TAAAAA	-27.8	(0.9)	†TAAATAG	-28.8	(1.7)	---	---	---	---	
41	6e-18	d11	AAGAAG	96.5	(55.5)	†GAAGGT	54.4	(38.3)	---	---	---	---	
42	2e-17	TRIM63	TCACTT	-1.9	(0.5)	CACTTC	-1.0	(0.3)	---	---	---	---	
43	3e-17	d12	AGTGCT	-529.4	(165.1)	---	---	---	---	---	---	---	
44	8e-17	TBP	AAAAGC	-26.9	(0.9)	ATATAAARGC	-29.9	(1.7)	---	---	---	---	
45	1e-16	Inr	CAGTGC	-1.0	(0.2)	---	---	---	---	---	---	---	
46	2e-16	d13	GGACCC	78.7	(27.8)	GGACCC	102.1	(46.4)	---	---	---	---	
47	3e-16	d14	GAGCCG	37.7	(36.2)	---	---	---	---	---	---	---	
48	6e-16	PDX1	GTCATT	-3.0	(0.5)	---	---	---	---	---	---	---	

Table 3.1: Location-specific motifs in human promoters. The location (μ) and width (σ) of overrepresentation are given to the right of each cluster. p -values given on the left pertain to the most significant 6-mer within the cluster. The third column shows factor names binding to the known regulatory elements in TRANSFAC [80]; putatively novel motifs are labeled d1-d14. Motifs found to exhibit locational specificity in mouse promoters are given to the right of the human analysis results. The right columns show comparisons to previous studies using the ‘sliding window method’ [38, 128, 145, 146]. Asterisks denote matches to non-redundant consensus motifs produced by these studies after kmer clustering; only motifs predicted to be enriched at approximately the same location were considered matches. All sequence matches to TRANSFAC, mouse motif predictions, and those of previous studies were conducted using STAMP [77] (E-value threshold: $E < 1e-6$).

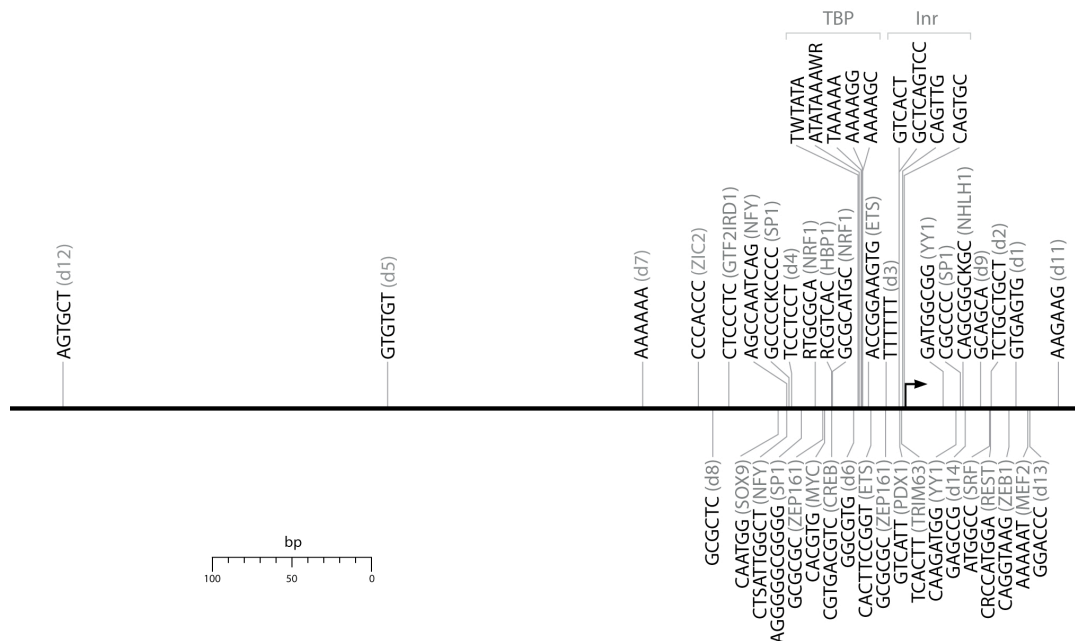


FIGURE 3.3: Location of overrepresentation for the 48 MLF motif predictions within human promoters.

The locational biases of these motifs may therefore reflect a functional role for each motif at these locations. The MLFs of the novel reverse complement motifs d2 and d9 are shown at the bottom of Figure 3.4; each orientation of this putatively novel regulatory element exhibits overrepresentation at the same location downstream of the TSS.

3.3 Many location-specific motifs are shared between human and mouse

We tested whether location-specific motifs found in human promoters would also show locational overrepresentation within mouse promoters. We conducted a second comprehensive MLF analysis using a sequence data set of 18,354 non-redundant mouse promoters in RefSeq [76, 101, 132] across the comprehensive list of all 6mer motifs. We then compared the motif predictions between the two species according to sequence similarity as well as the location of overrepresentation.

Our analysis predicted a total of 49 consensus motifs to exhibit location-specific overrepresentation within mouse promoters. Comparisons of these results to those of

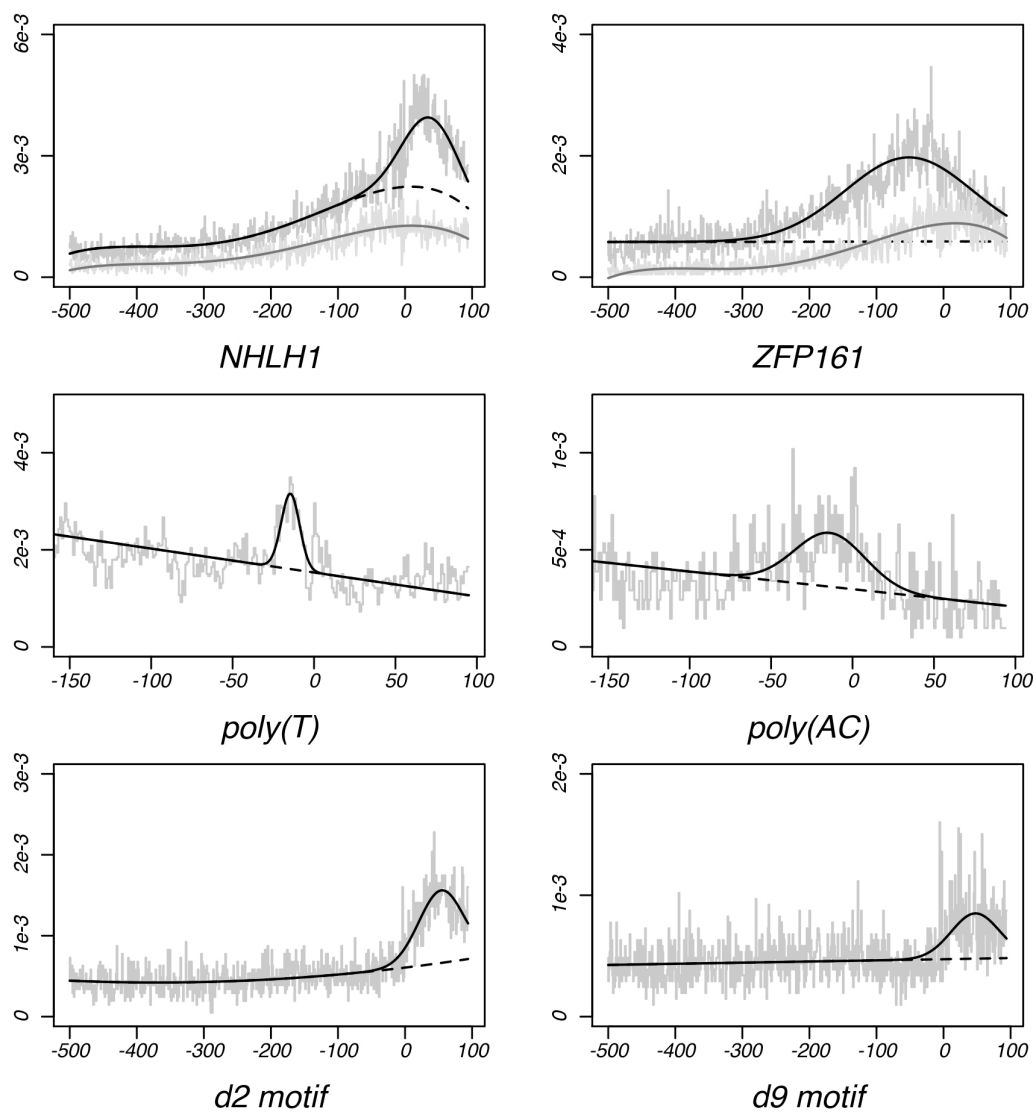


FIGURE 3.4: MLFs of six motifs exhibiting location-specific overrepresentation. Top: MLFs for the GC-rich NHLH1 binding site and the ZFP161 binding site. Each plot shows results for both real (black) and simulated (gray) data sets. Dashed lines denote the background functions $C(x)$. Middle: MLFs of the novel poly(T) 5mer d3 and (AC)₃ motif d10. Bottom: MLFs for the GCT-repeat motif d2 and its reverse complement d9. Each shows significant amounts of locational overrepresentation ~ 50 bp after the TSS.

the human promoter analysis showed a very significant amount of overlap between the motif predictions across the two species. We found that 36 (75%) of the motif clusters identified in the human data set matched location-specific motifs detected in the mouse promoters (Table 3.1). Such a significant overlap provides confidence in our new motif predictions, as these motifs were predicted during independent analyses using data from two highly divergent species. In addition, the location of overrepresentation for our motif predictions was often found to be highly conserved between the two lineages. Many motifs with well-documented locational specificity were found overrepresented at very similar locations across the two species, particularly the TBP, SP1, NRF1, and CREB binding sites. This was also found to be the case for many of our novel motif predictions. For instance, the novel d1 and d2 motifs exhibited overrepresentation peaks whose location differed by only 1 bp across the two lineages.

3.4 Study comparisons highlight differences in methodologies to previous studies

3.4.1 *The ‘sliding window’ approach*

Several previous studies have analyzed locational specificity of potential regulatory motifs within the promoter [38, 119, 128, 145, 146]. Most previous analyses, with one exception [119], have used the ‘sliding window’ approach. In this approach, the promoter region is divided into several discrete bins of pre-determined width (e.g., 20-25 bp), and locational specificity is then predicted by comparing the number of motif occurrences in each window to a background frequency of occurrence. A previous study conducted by FitzGerald et al [38] used the sliding window approach, considering motif occurrences within separate windows of 20 bp. FitzGerald et al predicted a total of 156 8mers to exhibit locational specificity prior to clustering. A direct comparison of our results to those of FitzGerald et al showed that 97% of the 8mers predicted by FitzGerald et al matched one of our predicted 6-mers (Table 3.2). We also found that 85% of our individual 6-mers matched an 8mer prediction

Positionally enriched motifs: Study comparisons					
		Fitz	Thara	Vard	MPF
Predictions	Number	156	1226	168	166
	kmer length	8	8	7	6
GC content	Fraction	63%	69%	60%	60%
GC rich	Number	28	387	19	48
	Expected	6	43	11	18
	(Numb / Exp)	4.7	9.0	1.7	2.7
AT rich	Number	3	39	0	16
	Expected	6	43	11	18
	(Numb / Exp)	0.5	0.9	0.0	0.9
FitzGerald	Matches	--	521	60	141
	Fraction		42%	36%	85%
Tharakaraman	Matches	149	--	101	156
	Fraction	95%		60%	94%
Vardhanabhuti	Matches	125	507	--	103
	Fraction	80%	41%		62%
MPF	Matches	151	1004	84	--
	Fraction	97%	82%	50%	

Table 3.2: Location-specific kmers are compared between studies conducted by FitzGerald *et al.* [38], Tharakaraman *et al.* [119], and Vardhanabhuti *et al.* [128] as well as the MPF model. The total number of (unclustered) kmer predictions are shown in the top row. The number of GC and AT rich motif predictions (those composed of G/C or A/T consensus sites at all but one site) are shown below, along with the expected number and the ratio actual/expected. Bottom rows show the amount of overlap between predictions across the four studies. Overlapping predictions were determined by considering all consensus sites of the predicted motifs, allowing for any offset such that at most one consensus site of the smaller motif was not aligned to the larger kmer. For instance, 149 (95%) of the 156 motif predictions made by FitzGerald *et al.* matched a prediction made by Tharakaraman *et al.* Note that the number of matches is not symmetrical, since a single kmer may match more than one other motif prediction.

made by FitzGerald *et al.* However, the vast majority of these matches were to redundant motifs that had been grouped according to sequence similarity during the clustering analysis. There were also cases in which distinct 6-mers found within different cluster groups matched a single 8mer predicted by FitzGerald *et al.* For instance, one of the G-rich 8mers predicted by FitzGerald *et al.* matched eight of our predicted 6-mers, although this group of 6-mers included representatives from three different motif clusters. These 6-mers clearly represented distinct regulatory elements, as their enrichment was found at significantly different locations within the regulatory region.

Thus, we compared the non-redundant consensus sequences produced by both

studies after clustering. Motif clustering conducted by FitzGerald et al resulted in nine non-redundant motif clusters. Eight of these clusters overlapped with one of our consensus motif cluster predictions, while our model attributed the putative locational bias of the remaining cluster to dinucleotide fluctuations within the promoter. In contrast, less than half of our consensus sequences were detected by FitzGerald et al. Table 3.1 contains comparisons between our regulatory motif predictions to those of FitzGerald et al as well as three other studies providing non-redundant motifs with locational overrepresentation [128, 145, 146]. We found that many of our motifs predicted with wider ranges of locational specificity could not be detected using the sliding window approach. Our approach was also found to increase sensitivity to location-specific overrepresentation occurring at very precise locations. For instance, FitzGerald et al, in addition to the three other studies included in Table 3.1, could not easily detect the well-known Inr sequence element. Inr sequence has been previously characterized by the consensus motif YYAnWYY [148]. This element is known to function specifically at a single nucleotide site at the start of transcription [98, 148], and therefore it is difficult to detect using low resolution approaches. Out of 156 8mer predictions made by FitzGerald et al, none included the YYAnW 5mer overrepresented at the TSS. In contrast, our model identified seven 5mers matching this consensus with significant enrichment at the start of transcription ($p < 1e-15$). The most common version of this motif was TCAGT, which was found overrepresented at the TSS more than seven and a half times over the background frequency ($p = 6e-48$). Despite the highly significant amount of locational specificity exhibited by this motif, none of the studies using the sliding window approach detected any motifs containing this 5mer [38, 128, 145, 146]. Figure 3.5 shows the occurrence data of this motif using 20 bp windows and using single-site resolution; we note the significant decrease of the signal when considering the data using windows of 20 bp.

3.4.2 Comparisons to Tharakaraman et al.

Tharakaraman et al [119] also scanned for locational specificity within human promoters. However, their methodology allowed for varying window sizes, improving

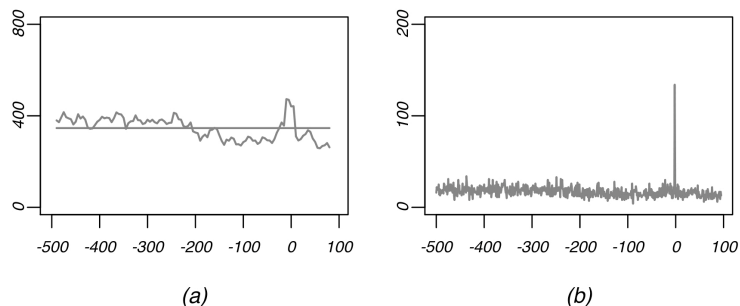


FIGURE 3.5: Occurrence frequency of the functional Inr sequence 5mer TCAGT. The contrast is shown between occurrence data using (a) 20 bp windows and (b) single-site resolution.

sensitivity for locational overrepresentation considerably. Tharakaraman et al predicted 1226 unclustered 8mers to exhibit locational specificity within the promoter. Despite such a large number of predictions made by Tharakaraman et al, 82% of their predicted motifs overlapped with our results (Table 3.2). However, their model assumed a uniform background frequency of occurrence across the promoter. Since GC mono- and di-nucleotide composition rises substantially near the start of transcription [38], about a third of the 8mers predicted by Tharakaraman et al were highly GC-rich, containing at least seven out of eight G/C consensus sites. This is nine times more than what would be expected from a random selection of 8mers. In contrast, the number of GC-rich motifs predicted during our analysis is only 2.7 times higher than would be expected by chance.

As many GC-rich motifs do play functional roles in gene regulation, we looked to determine whether these GC-rich 8mers do, in fact, comprise true regulatory elements. To assess the validity of these GC-rich predictions, we compared the predictions made by Tharakaraman et al to known protein-binding sites found in humans. We found that, among the GC-rich predictions overlapping our results, over half matched human binding elements found in the TRANSFAC database [80]. This represents a significant enrichment of known regulatory elements, as only about a third of all GC-rich 8mers match human binding sites in TRANSFAC. However, among the GC-rich predictions that did not overlap our results, only 19% matched

known human binding sites. This is significantly less than would be expected by chance given a random selection of 8mers. Although this evidence is not necessarily conclusive, we would still expect some amount of enrichment for known regulatory elements in this list of predictions. Thus, it is likely that a number of these predictions are simply the result of the rise in GC content near the TSS, rather than true regulatory elements.

3.4.3 Comparisons to Vardhanabhuti et al.

In contrast to the analysis of Tharakaraman et al, Vardhanabhuti et al [128] controlled for changes in basepair composition across the promoter. In this analysis, the observed number of occurrences of a given motif was compared to an expected number of occurrences in each window of 20 bp. The expected frequency was estimated separately within each individual window by considering occurrence data of other motifs with identical basepair composition. That is, occurrence data was obtained for motifs whose columns were ‘permuted’ from the original motif, thus conserving base composition. The observed occurrences of these permuted motifs were then used to determine the expected frequency of occurrence in each individual window; both the ‘observed’ and ‘expected’ frequencies were thus unique to each window. Vardhanabhuti et al first scanned for locational overrepresentation using known transcription factor binding sites in TRANSFAC, and subsequent analyses predicted location-specific overrepresentation across all (novel) 7mer motifs filtered for known binding sites in TRANSFAC. Although these latter predictions were presented as novel motifs, about a third of these 7mers matched known regulatory elements. The top fifteen ‘novel’ motifs predicted by Vardhanabhuti et al are given in Table 3.3, along with the matching TFBS found in TRANSFAC.

Between-studies comparisons showed consistently less overlap between the results of Vardhanabhuti et al and those of other studies, including the one presented here (Table 3.2). It is likely that these differences can be explained by the methodology used to estimate the background frequency of occurrences. For instance, the occurrence frequency of a motif rich in a single nucleotide type will not be signif-

Rank	Consensus	TF	TFBS	E-value
1	AGATGGC	NF- μ E1	AGATGGC	2e-11
2	ATTGGCT	alpha-CP1 NFY	ATTGGCT ATTGGYT	3e-10 8e-8
3	CCGACAT	--	--	--
4	CACTTCC	GABP ETS	CnCTTCC nACTTCC	1e-9 1e-8
5	GGTGAGT	--	--	--
6	AGCCAAT	alpha-CP1 NFY	AGCCAAT ARCCAAT	3e-10 8e-8
7	GCGGGGC	SP1	GCGGGGn	7e-7
8	GGAAGTG	GABP ETS	GGAAGTG GGAAGTn	1e-10 1e-8
9	CCATTGG	[†] alpha-CP1	TCATTGG	1e-6
10	GTCAATC	COMP1	GTCAATC	2e-7
11	GACGTAA	CREB	GACGTMW	1e-9
12	GAAGATG	[†] YY1	nAAAnATG	1e-6
13	CTGATTG	NFY	CTGATTG	3e-9
14	TATAAGG	SRF [†] TBP	WATAAGG TATAAAAn	1e-7 7e-6
15	ATGGCGG	E2F	TTSTCGG	3e-7

Table 3.3: ‘Novel’ motif predictions made by Vardhanabhuti et al [128]. The top fifteen 7mer predictions presented by Vardhanabhuti et al as novel motif predictions are shown in the 2nd column, along with the TFBS consensus sequence within TRANSFAC (4th column) as well as the regulatory *trans*-factor binding protein (3rd column). E-values representing the statistical significance of the matching sequences (as determined by STAMP [77]) are given in the last column.

icantly different after permuting its columns, as the motif consensus itself will not be significantly changed. In particular, mono-nucleotide repeats are impossible to detect. As a result, sensitivity to many biologically relevant signals is decreased significantly. Vardhanabhuti et al note within their study that their methodology predicts enrichment of the well-known TBP binding element (TATA-box) at a location that differs from where it is known to function. This motif was predicted by Vardhanabhuti et al to be overrepresented 45 bp prior to the TSS, although it is known to function at a very specific location 30 bp upstream of the TSS [148]. The authors attribute this discrepancy to an increase of A/T nucleotide composition at this location, increasing the ‘expected’ number of occurrences within this window and therefore decreasing the observed/expected ratio. However, the increase of A/T nucleotide composition at this location is simply a result of the overrepresentation

of the A/T rich TBP binding site itself. This raises the concern that correcting for basepair composition in a location-specific manner can cause failure to detect real biological signals, as the signal itself can be incorporated into the background (expected) frequency. The method presented here effectively circumvents this problem, as the background frequency is modeled in a continuous fashion. Significant changes in the expected frequency caused by real biological signals remain outlier points after fitting the background model to the data (Figure 3.1b). We note that in the case of the TATA-box, the MLF method predicted overrepresentation at the correct location 30 bp prior to the TSS at a high level of confidence.

Motif Relational Functions

Transcription is not driven by single protein factors acting in isolation, but is instead controlled by multiple regulatory elements acting in coordination [4, 10, 33, 144]. Knowledge of regulatory element interactions is therefore essential to understand the mechanisms driving gene regulation. Since protein-protein interactions are inherently structure-specific, it is logical to expect that motifs binding interacting proteins preferentially co-occur non-randomly with respect to each other. It has been previously shown that many regulatory motifs do exhibit spatial preferences relative to each other [2, 5, 43, 51, 110, 128, 129]. However, most studies analyzing inter-motif distance preferences have generally been limited to discovering spatial relationships between either known protein binding sites or motifs predicted using standard motif finding methods.

Since the MPF model is designed to detect positional specificity of a motif relative to a given reference point, it can also provide a means by which to study distance preferences between pairs of motifs. While the model described in the previous chapter can predict distance preferences of a single motif relative to the TSS, in this chapter, we assess spatial preferences between pairs of potential regulatory

elements. Thus, we use the MPF model to measure distance preferences of a given motif relative to a known occurrence of a second motif. Our approach is to comprehensively search for spatial inter-dependencies across all possible (5mer) motifs-pairs on a genome-wide scale. This allows us to predict motif-pairs binding interacting regulatory proteins *de novo*, without prior knowledge regarding the sequences of the individual regulatory elements. We denote this second version of the MPF model as the ‘motif relational function’ (MRF) model.

4.1 Motif relational functions (MRFs) detect spatial biases between motif-pairs

4.1.1 Overview of the MRF model

The MRF model represents a simple extension of the MLF model. In this extended model, we choose the reference point to be a known occurrence of a particular motif (v). For a second motif (w), an MRF ($f_{w|v}(x)$) then represents the underlying occurrence frequency of w relative to v . In this case, the position $x = 0$ represents an occurrence of motif v . Positive values for x then represent the distance of w downstream of v , while negative x -values give the distance of w upstream of v . The MRF $f_{w|v}(x)$ then represents a conditional probability of occurrence for w at x , assuming the motif v occurs at the location $x = 0$ within the same sequence (Appendix B.1).

Modeling inter-motif distances preferences using an MRF is more complex than the MLF model, which assumes that overrepresentation can only occur at a single location within the promoter. It has previously been shown the pairs of DNA sequence elements can bind interacting transcription factors when separated by one of several preferred distances, with protein-protein interactions occurring at phased intervals across the DNA [53, 69, 135, 144]. Such periodic distributions have been associated with DNA sequence features attributed to the structural conformation of the nucleosome [53] as well as the histone complex or the winding of the DNA

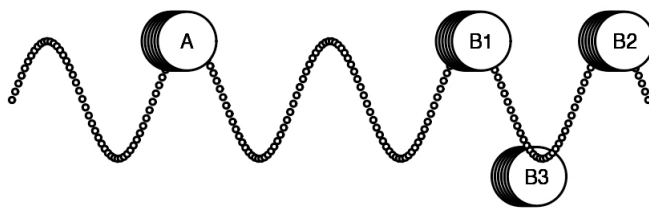


FIGURE 4.1: Functional motif-pair inter-relationships. Proteins must often be positioned in a particular orientation with respect to the DNA molecule to induce potential interactions [69, 135, 144]. Interactions between protein A and protein B occur when the latter is positioned at B1. The same interaction frequently occurs one turn of the double-helix away from B1 (i.e., at B2), since the orientation of protein B is consistent relative to the turn of the DNA molecule. However, the interaction cannot occur when protein B is bound at B3 due to its inconsistent orientation. The distance between factors A and B is determined by the size of the proteins and is therefore unique between different transcription factor-pairs. In contrast, phasing intervals (i.e., the distance between B1 and B2) remains relatively consistent across factor-pairs, as they correspond approximately to the number of nucleotides in a turn of the DNA double-helix.

double-helix [69, 135, 144]. This scheme is shown in Figure 4.1, which illustrates a potential preference for protein-protein interactions to occur in a specific orientation in relation to the turn of the double-helix.

Figure 4.2 shows two MRFs which were both generated by motif-pairs that bind transcription factors with known interactions [6, 73], namely, the NFY-NFY and NFY-SP1 binding motif-pairs. Motif-pairs were often found to co-occur preferentially at multiple distances, with intervals separating preferred distances corresponding approximately to the turn of the DNA double-helix.

4.1.2 The MRF model incorporates phased distance preferences between motifs

In order to capture the periodic nature of inter-motif distance preferences, we allow each MRF to exhibit multiple preferred distances separating each pair of motifs. This was done by extending the signal function model $H(x)$, as described in the previous chapter, to allow for multiple overrepresentation peaks. Thus, the the MRF signal

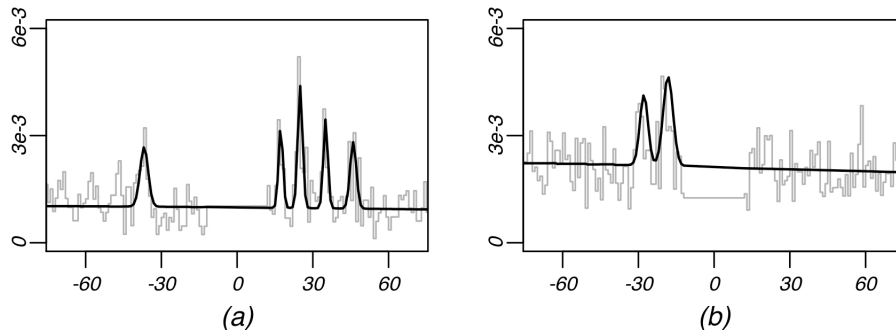


FIGURE 4.2: MRFs of two motif-pairs binding interacting transcription factors [6, 73]. A known occurrence of the reverse-strand NFY binding site defines the position $x = 0$; x -axis values denote the position of the (a) plus-strand NFY binding site and the (b) minus-strand SP1 binding site. y -axis values show the frequency of occurrence of these partner motifs. Each motif-pair exhibits more than one preferred distance between motifs, with intervals between peaks being ~ 8 -10 bp. This is consistent with the scheme illustrated in Figure 4.1, where the location of the peaks represent the positions of B1 and B2, and the position $x = 0$ corresponds to the position of factor A.

function takes on the form:

$$H(x) = \sum_{j=1}^M a_j \cdot \exp \left[-\frac{(x - \mu_j)^2}{2\sigma_j^2} \right] \quad (4.1)$$

This function $H(x)$ contains M overrepresentation peaks, each representing a single distance at which the two motifs prefer to occur. Note that this is a generalization of the signal function used during MLF analyses, which only allows for one peak (see Eq 3.1). Here, the number of overrepresentation peaks M is not pre-defined but is instead estimated separately for each individual motif-pair. Our model is designed to predict overrepresentation peaks on an individual basis, continuing to add overrepresentation peaks until all statistically significant peaks have been incorporated into the model.

4.1.3 *Phased intervals between preferred inter-motif distances corresponds to the turn of the DNA double-helix*

In order to determine whether phasing intervals between overrepresentation peaks are consistent across different motif-pairs, we determined inter-motif distance preferences across all possible pairwise combinations of 5-mer motifs. Our comprehensive MRF analysis was thus conducted across each pair of 5-mers, estimating the MRF produced from each motif-pair. For each motif-pair that produced at least 2 overrepresentation peaks, we determined the separation distance between each pair of overrepresentation peaks. In order to quantify the phasing of inter-motif distance preferences, we define a ‘peak separation value’, x_{sep} , to be the distance between any pair of overrepresentation peaks within the same motif-pair MRF. We set this value to be $x_{sep} = |\mu_i - \mu_j|$ for any two peaks i and j within the same MRF signal function, i.e., two different Gaussian terms in $H(x)$ as given in Eq 4.1. We controlled for peaks potentially representing random outlier points by filtering peaks corresponding to a single-site location. We also filtered double motif occurrences separated by less than 20 bp in order to remove spurious peak phasing caused by repeat sequences and same-sequence dyads.

Enumerating peak separation values across all MRFs containing at least two overrepresentation peaks, we found that 619 MRFs exhibited x_{sep} -values ranging between 7.5-9.5 bp (Figure 4.3). This is more than twice the expected number (i.e., the average number of MRFs producing x_{sep} -values within any other 2 bp range). While these values do not correspond precisely to the number of nucleotides in a turn of the double-helix (~ 10.5 bp), it is possible that this deviation can be explained by distortions of the DNA caused by protein binding or by other similar mechanisms (see Section 4.4 for discussion).

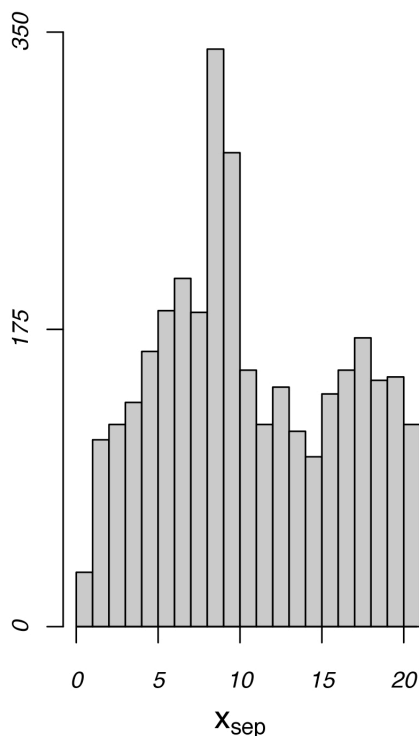


FIGURE 4.3: MRF peak separation distributions. The distribution of peak separation (x_{sep}) values shown are produced by enumerating all peak-pairs across all possible 5-mer motif-pairs. Peak separation values are defined as $x_{sep} = |\mu_i - \mu_j|$ between overrepresentation peaks i and j within the same MRF; this value is analogous to the distance between B1 and B2 in Figure 4.1. Note the strong concentration of x_{sep} -values close to ~ 8 -9 bp.

4.2 Predicting regulatory motifs using peak separation values

4.2.1 Several motifs exhibit consistent peak separation values

Given the high concentration of x_{sep} -values around 8-10 bp, we hypothesized that motifs producing consistent x_{sep} -values corresponding to the turn of the double-helix would act as protein binding sites. We chose to use a stringent criterion for motif-pair predictions by specifically focusing on motifs exhibiting significant concentrations of x_{sep} -values, whose consistency was unlikely to be due to chance. Thus, for each individual (fixed) 5-mer motif, we calculated the distribution of peak phasing intervals across all MRFs produced from this fixed motif and one of the possible (variable) motif partners. x_{sep} -values were then accumulated across all variable motif partners.

Motifs with consistent phasing intervals					
Rank	Consensus	TF	x_{sep}	p	Partners
1	TTTGTA	y1	9	7e-26	19
2	ATTTTT	MADS (-)	8	3e-21	24
3	AAAAAT	MADS (+)	8	4e-19	16
4	GCATGC	NRF1	9	7e-19	23
5	ATTGC	y2	8	3e-12	8
6	TCTTG	EVI1	9	1e-11	7
7	GAGCT	y3	10	2e-10	7
8	ATTGG	NFY (-)	10	4e-8	5
9	CCAAT	NFY (+)	10	1e-7	3

Table 4.1: Consensus motifs with consistent phasing intervals. Phasing intervals (x_{sep} -values) were considered across all MRFs produced by the (fixed) motifs shown above and one of the possible 5-mer partners. x_{sep} -values denote the interval between preferred inter-motif distances (i.e., overrepresentation peaks) within the same MRF for a pair of motifs ($x_{sep} = |\mu_i - \mu_j|$). x_{sep} -value concentrations were determined across all 2 bp intervals centered around 8-10 bp. p -values (fifth column) correspond to the significance of this concentration for the top-ranking 5-mer in each cluster. Transcription factors binding to known motif in TRANSFAC [80] are shown in the third column (STAMP [77] E-value threshold: 1e-5); novel regulatory element predictions are label y1-y3. The number of predicted partner clusters is given in the right column.

Fixed motifs producing a significant x_{sep} -value concentration within one of the 2 bp windows centered at $x_{sep} \approx 8, 9, 10$ were then predicted to be functional.

After correcting for multiple hypothesis testing, thirteen 5-mers were found to have significant x_{sep} concentrations within one of these regions ($p < 1e-5$). Clustering these 5-mers according to sequence similarity and their corresponding x_{sep} distributions produced nine-consensus motifs (Table 4.1). Six of the nine consensus sequences matched known TFBSs in TRANSFAC, namely the NRF1, NFY, EVI1, and MADS-box protein family binding sites. The NFY and MADS-box protein family binding motifs were predicted on both strands, while the NRF1 binding sequence was palindromic. Three additional motifs, denoted as y1-y3, did not match any known binding sequences in TRANSFAC and therefore represent novel *cis*-regulatory element candidates. Figure 4.4 shows the x_{sep} -distribution for the highest-ranking 5-mer in four of the predicted motif clusters. These include the reverse-strand MADS-box protein family binding site, the NRF1 binding site, the novel y1 motif, and the reverse-strand

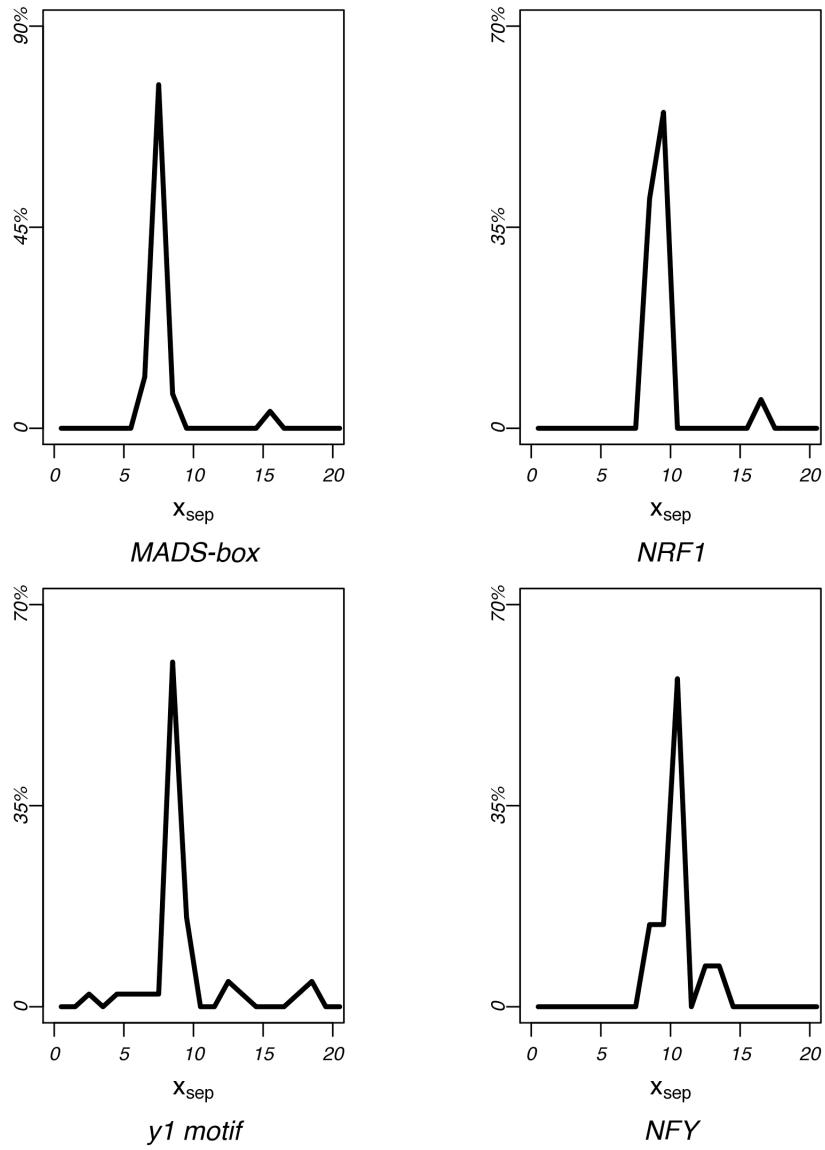


FIGURE 4.4: MRF peak separation concentrations. Peak separation distributions are shown for four motifs with significant x_{sep} -value concentrations. Each panel shows the x_{sep} distributions for the most significant 5-mer of the reverse-strand MADS-box family binding motif, the NRF1 binding motif, the reverse-strand NFY binding motif and the novel regulatory element prediction y1.

NFY binding site. Note the highly significant concentration of x_{sep} -values around \sim 8-10 bp for each motif.

4.2.2 Periodic phasing of inter-motif distance preferences detects known and novel regulatory element relationships

We extended the analysis in order to predict binding site partners for each of the motifs exhibiting significant x_{sep} -value concentrations. Each fixed motif predicted during the previous MRF analysis was paired with multiple partner motifs by considering each individual fixed/partner motif-pair MRF. A 5-mer was predicted to pair with the fixed motif if the motif-pair produced phased distance preferences corresponding to the fixed motif's x_{sep} concentration. The predicted partner motifs were then clustered according to sequence similarity as well as the location of their over-representation peaks. This procedure produced a total of 112 motif partner clusters pairing with one of the nine fixed motifs predicted in the previous analysis.

Partner motif predictions for the NFY and MADS-box protein family binding sites are given in Tables 4.2 and 4.3, respectively. Only a few motif clusters were predicted to pair with the NFY binding motifs; each of the partner clusters corresponded to either the NFY or the SP1 binding sequences. Both factors are known to have direct interactions with NFY [6, 73].

The MADS-box protein family binding sites were predicted to pair with more partner motifs than the NFY binding element. The MADS-box family consensus sequences predicted during the analysis bind both the myocyte enhancer factor 2 (MEF2) and the serum response factor (SRF) [50]. These two factors are known to be involved in complex extra-cellular signaling pathways, playing multiple roles involving cell differentiation and development [23, 82, 99]. Both MEF2 and SRF regulate gene expression through the recruitment of multiple accessory co-factors whose presence or absence within the complex cause differential expression of their

NFY binding site partners		
TF partner	Forward strand	Reverse strand
NFY (+) (-)	CCAAT	GCCAATC
	.GATTGGC CGATT...	ATTGG
SP1 (+) (-)	GGCGG	GGGCGG
	CCGCC	

Table 4.2: Motif partners for the NFY binding element. Partner motifs for the forward and reverse strand NFY binding motif are shown in the second and third columns, respectively. Each partner motif binds either the NFY or SP1 factors (left column). Both NFY-NFY and NFY-SP1 factor-pairs exhibit known interactions [6, 73].

MADS-box motif partners							
Forward strand: AAAAAT				Reverse strand: ATTTT			
	Consensus	TF	RC		Consensus	TF	RC
1	AGACC			1	GAACTCCT	NR1I2	
2	CAGCTAC	TOPORS		2	AGCCT		R1
3	AGGCTG		r1	3	AGTGC	HMX3	R8
4	CAGCC			4	ATCCG		
5	CCTGTA	AR		5	ATGTT		
6	CGCCA	E2F1		6	CACCA	NFY (*)	R7
7	CTACTC			7	CCACG	ATF6 (*)	
8	GCTGAG	NFE2	r2	8	CCCAA	IKZF1	
9	GAACC			9	CCTCC	MAZ	R3
10	GGCAGG	TCF3 (*)		10	CCTGA		
11	GGAGG	MAZ	r3	11	TCGAAC	XBP1	
12	GTTTG		r4	12	GCTGGGACA	PITX2	
13	TGTAATCCCA	CEBP (*)	r5	13	GATCC		
14	GTGGC		r6	14	GGATTACA	CEBP (*)	R5
15	TGGTG	NFY (*)	r7	15	GCCAC		R6
16	GCACT	HMX3	r8	16	GGGTTT	TERF2IP	
				17	TCAAG	NKX2	
				18	TGACC	ESR1	
				19	TGATC		
				20	AGCCA	PCBP2	
				21	CAACC		R4
				22	CTCGG	ZNF569	
				23	TCAGC	NFE2	R2
				24	TGCCT		

Table 4.3: MADS-box family binding site partners. Transcription factors binding to the known *cis*-regulatory elements in TRANSFAC [80] are shown in the third columns. Binding factors with known direct interactions to SRF, a MADS-box family member, are labeled with asterisks. Reverse complements across opposing strands of the MADS-box fixed motif, as determined by STAMP (E-value threshold 1e-5) are labeled r1-r8 and R1-R8 (e.g., r1 is the reverse complement of R1, etc.).

target genes [41, 42, 134], and therefore we would expect a large number of partner motifs to pair with their binding elements.

Sixteen and 24 partner clusters were predicted to pair with the forward- and reverse-strand MADS-box motif, respectively. We found that partner motifs pairing with the MADS-box binding site were frequently predicted in both orientations. Eight reverse complement-pairs were predicted to occur across opposing strands of the MADS-box binding motif (Table 4.3). The mutual directionality between the MADS-box binding motif and its partner motifs was highly conserved, with each individual strand of the partner motif pairing with one, but not both, orientations of the MADS-box binding motif.

A total of 24 (60%) of the partner clusters were found to match known protein factor binding sequences in TRANSFAC, comprising a total of 19 known regulatory elements. Several of these regulatory elements are known to bind proteins with direct interactions to SRF, including the binding motifs of TCF3 [41], CEBP [42], NFY [147], and ATF [152]. Twelve motif partners bind proteins known to be involved in either signal transduction pathways or developmental processes. Three such factors belong to the homeobox family, whose members play a crucial role in early development [74, 90, 136]. Many of the remaining partner motifs may play unknown functional roles in concert with one of the MADS-box protein factors.

4.3 Uni-modal versus multi-modal approaches

In our application of the MRF model, we explicitly use multi-modal characteristics of the inter-motif distance frequencies as a criterion for functional motif relationships. This approach is inherently different from previous uni-modal models, which have generally relied on the sliding window method or maximum-distance approaches [2, 43, 51, 110, 128, 129]. While uni-modal approaches have had some success in predicting motif inter-dependencies, our data suggest that high-scale resolution is

often necessary to detect spatial relationships between motifs. We have found that individual instances of spatial preferences are generally constrained to widths of only $\sim 2-3$ bp, and that single overrepresentation peaks often exhibit only minimal amounts of significance. However, despite the subtlety of each individual instance of spatial preference between motif-pairs, overall trends in the phasing intervals between preferred inter-motif distances are highly significant.

Our model was intentionally designed to account for spurious overrepresentation peaks by considering peak separation distances collectively across a comprehensive list of all pairs of 5-mers. We explicitly focus on motifs with distinguishing phasing intervals between preferred distances, and whose consistency was unlikely to be due to chance. This represents only one particular application of the MRF model, and as more knowledge is gained regarding *cis*-regulatory element relationships, other applications of this model may also prove effective.

4.4 Deviation of phasing interval values

As noted previously, we found that many motif-pairs exhibit multiple preferred separation distances with phasing intervals often found near the range of $\sim 8-9$ bp. This corresponds to approximately 2 bp less than the number of nucleotides in one turn of the DNA double helix (~ 10.5 bp). The deviation from the expected number of 10.5 bp is worth noting, particularly due to the robustness of the signal occurring across such a large number of motif-pairs. There are some possible explanations available from the literature. Structural analyses have shown that protein binding, and the binding of multi-protein complexes in particular, distort the conformation of the DNA [54, 92, 100, 123], thus affecting the helical characteristics of the DNA. Interestingly, the MADS-box family transcription factors in particular are known to bend the DNA molecule upon binding [134], largely explaining the consistent deviation of the phasing intervals for this binding motif from 10.5 to approximately 8

bp.

An alternative, although not necessarily exclusive, interpretation is that the occurring protein-protein interactions may either be stabilized by alterations of the DNA molecule or require them for collective binding. Selective binding of proteins to DNA involves not only sequence-specific elements within the DNA, but also topological characteristics of the DNA molecule [45, 63, 134]. This is known to be particularly true during the recruitment of multiple interacting proteins to the DNA [63, 134]. Thus, although the biological explanation of the observed pattern remains unclear, these results are not inconsistent with our current knowledge of protein-protein and protein-DNA interactions.

Location-specific *cis*-regulatory element evolution

Changes in DNA sequence elements that control gene expression can have broad impacts on species morphology [66, 144, 143]. A single transcription factor can sometimes regulate the expression of hundreds or even thousands of genes genome-wide, binding to commonly occurring DNA regulatory motifs in a sequence-specific manner [38]. Modifications within the preferred protein-binding sequences therefore involve a massive number of changes in order to preserve the set of target genes regulated by the corresponding trans-factor. As such changes must each occur independently across the genome both to and from specific nucleotides, it is often assumed that these DNA binding motifs are rarely modified over the course of evolution [44, 72, 128, 146].

In this work, we take a genome-wide approach to assess the prevalence and nature of regulatory motif modifications within vertebrates. We examine to what extent the genome can adapt to global evolutionary changes in *cis*-regulatory motifs, such as those illustrated in Figure 5.1. We show that a very substantial fraction of location-specific *cis*-regulatory motifs have experienced significant modifications over the course of evolution, even within relatively closely-related mammals. Fur-

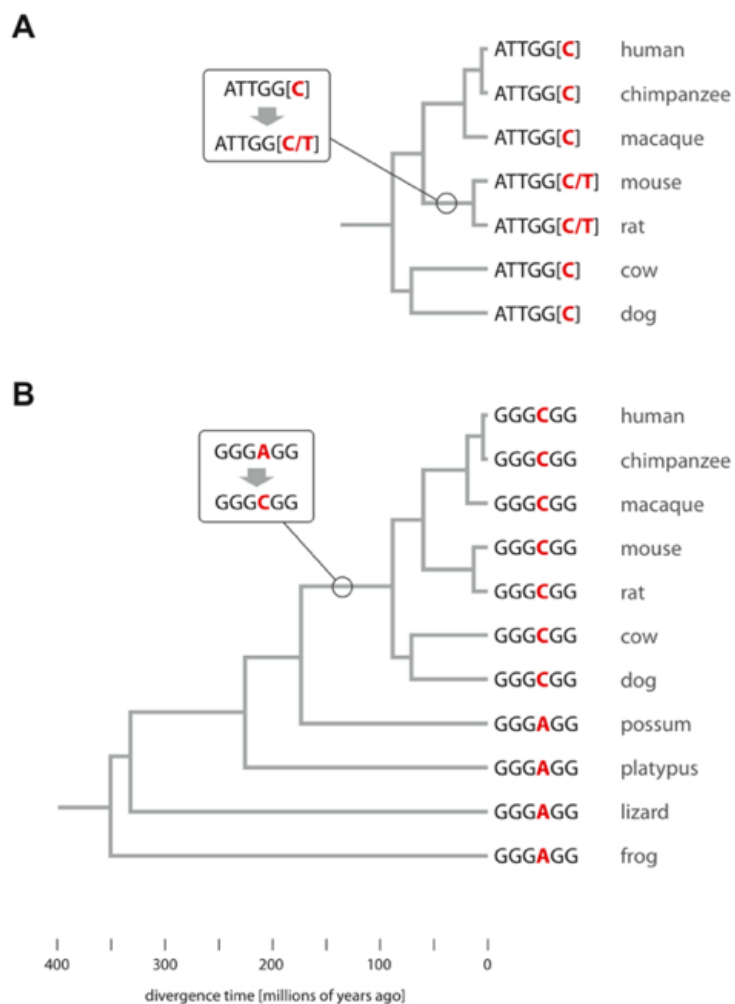


FIGURE 5.1: Examples of *cis*-regulatory motif modifications. Shown are lineage-specific changes within the (A) NFY binding site and the (B) SP1 binding site. Branch lengths are drawn to scale, with evolutionary divergence times as estimated in [47].

thermore, such changes are frequently found to occur at a very rapid rate, often surpassing the background rate of substitution. This work identifies a previously unrecognized mechanism driving genome sequence evolution, and provides important insights into the evolution of regulatory sequences in general.

5.1 Studying location-specific *cis*-regulatory element evolution

In our study, we focus upon evolutionary modifications occurring specifically within location-specific motifs. Location-specific overrepresentation is a convenient characteristic by which to study motif evolution. Rises in motif occurrence frequency suggest that the motif plays a functional role in gene regulation at that location within the promoter, while occurrences outside the region of preference are presumably less likely to perform the same regulatory function. For certain motifs this is known to be the case, such as with the commonly occurring TATA-box motif, which is found highly overrepresented ~ 25 -30 bp prior to the TSS, where it is known to function [79, 149]. In previous studies, it has been shown that the location of motif occurrence, i.e., occurrences within versus outside the region of overrepresentation, correlates with function and expression patterns of the target gene [128]. We find that the location-specific motifs do, in fact, evolve differently within their region of overrepresentation than outside this region; this is discussed in Section 5.1.4. Our methodology utilizes the ability to distinguish between functional and ‘background’ motif occurrences. Here, we compare evolutionary modifications within functional motif occurrences inside the region of overrepresentation to substitution frequencies within the less constrained ‘background’ motif occurrences outside this region.

5.1.1 Location-specific motifs are shared across closely and distantly related species

In Chapter 3, we discussed the high amount of overlap of location-specific regulatory elements between mouse and human. To test whether such overlap existed between mammals and non-mammalian vertebrates, we scanned for location-specific motifs in zebrafish [133] and compared the predictions to those within mouse. These comparisons showed that most motifs were shared across even between these highly diverged lineages. Table 5.1 shows a comparison of the 20 top-ranked location-specific motifs

Location-specific motifs in Mouse and Zebrafish										
#	TF	Mouse				Zebrafish				Overlap # Sites
		Consensus Motif	Top 6mer	Peak	Width	Consensus Motif	Top 6mer	Peak	Width	
1	SP1	KCCCCCKCCCM	CGGCC	-73	100	CCCCTCCY	CCTCCC	-67	100	94
2	TBP	TMTATAAARGc	TATAAA	-30	6	NSTATAAAAGc	TATAAA	-30	6	6
3	NFY	AGCCAATSAG	GCCAAT	-83	100	AGCCAATCA	GCCAAT	-88	100	95
4	CREB	GTSACGTGA	TGACGT	-44	100	CGTGACGTC	TGACGT	-49	100	95
5	SP1	GNGGGGGGCGK	GGGCGG	-63	100	GGGAGGGGG	GGGAGG	-76	100	87
6		GTGTGTG	TGTGTG	-440	100					
7	NFY	CTGATTGGY	ATTGGC	-79	100	CTGATTGGCT	GATTGG	-83	100	94
8		CRCCATGGMn	CCATGG	+52	100	ACATGGCT	CATGGC	+22	64	52
9		MAGGTRAGTG	GGTAAG	+71	100	GTAAGW	GTAAGT	+65	89	89
10	ETS	SCGGAAGTG	CGGAAG	-31	100	MGGAAGT	CGGAAG	-21	100	90
11	ERF2	CAGCGGCSGC	GGCGGC	+35	100					
12	HBP	RCGTCAC	ACGTCA	-47	100	CACGTG	CACGTG	-50	100	97
13	E2F	TGGCGG	TGGCGG	+26	54	TGGCGG	TGGCGG	+18	28	28
14		YGC GCGC	GCGCGC	-29	100	CGCGCGC	GCGCGC	-46	100	85
15	CREB	ACTTCCGG	TTCCGG	-20	74	WCTTCCT	ACTTCC	-31	100	74
16	NRF1	TGCGCA	TGCGCA	-59	100	GCATGCGCGT	ATGCGC	-46	100	87
17		TCTGCTGCT	GCTGCT	+58	100	GCTGCTGC	CTGCTG	+49	100	91
18	NF- μ E1	GRTGGC	GATGGC	+29	66	RATGGC	GATGGC	+16	30	30
19		AAAAAA	AAAAAA	-104	100	AAAAAA	AAAAAA	-93	100	89
20	YY1	ASATGG	AGATGG	+17	34	ACATGGCT	CATGGC	+22	54	34

Table 5.1: Comparisons between location-specific regulatory motifs within mouse and zebrafish. The left columns show the top 20 motifs exhibiting location-specific overrepresentation in mouse. The top-ranked 6mer in each cluster, center of the region of overrepresentation (μ), and the width of this region ($3 \cdot \sigma$) are shown to the right of each consensus motif. Matching motifs predicted for zebrafish (danRer5, Zv7 assembly [133]) are shown at the right. The number of overlapping nucleotide sites within the region of overrepresentation between the two species are given in the last column. Note the strong tendency for the location of overrepresentation to remain highly conserved even across these very distantly related species.

found in mouse to predicted location-specific motifs within zebrafish. Of these motifs, eighteen were found to overlap with location-specific motif predictions in zebrafish. The location of overrepresentation for these motifs was found highly conserved, with the center of the peak within 15 nucleotide sites in the majority of cases.

5.1.2 Determining functional motif co-occurrences across species

In order to study the nature of evolution within *cis*-regulatory elements, we consider motif occurrences that target orthologous genes across species. Within a given motif, we search for potential evolutionary modifications at each individual consensus site separately. Considering a single chosen consensus site within a given motif, our goal is to determine the presence or absence of cross-species differences in nucleotide frequencies at the given site. In order to assess cross-species differences in nucleotide

		Human			
		attggA	attggC	attggG	attggT
Mouse	attggA	43	9	14	3
	attggC	6	341	9	38
	attggG	8	3	40	1
	attggT	2	80	5	104

FIGURE 5.2: Cross-species nucleotide co-occurrence data across mouse and human at the 6th site of the NFY binding element (motif ATTGGn). Individual elements within each of the matrices represent the number of motif co-occurrences across species within the region of overrepresentation.

preferences at the chosen consensus site, any 6 bp element matching the motif at the five non-chosen sites was considered to be an occurrence of the motif, while we allowed for one possible mismatch at the chosen site. Co-occurrences of the motif targeting orthologous genes within the region of overrepresentation were then determined. Each cross-species co-occurrence then contains a specific nucleotide at the chosen (mismatch) site in each of the two species.

In this way, we determine the number of nucleotide co-occurrences for each pairwise combination of nucleotides at the given consensus site across the entire set of orthologous promoters. These co-occurrence counts can be illustrated using a 4x4 matrix, such as those provided in Figure 5.2. Note that each consensus site produces a single matrix, and therefore six different co-occurrence matrices were generated for each motif, one for each site. In order to avoid ambiguity regarding multiple motif occurrences within the same gene of the same species, we filter double motif occurrences found within 150 bp of each other. We then use the nucleotide co-occurrence matrices to search for differences in nucleotide preferences across species.

5.1.3 Determining background motif co-occurrences

In order to determine the statistical significance of lineage-specific modifications within each motif, we compared nucleotide occurrence frequencies within the re-

gion of overrepresentation to those found in background motif occurrences within intergenic sequences. Each intergenic sequence comprised the 1 kb region starting 2 kb upstream of a known TSS. Orthologous pairs of intergenic sequences were determined using the multiz sequence alignments [85]. We then determined orthologous occurrences of the motif in a similar manner as those within the region of overrepresentation. Specifically, given a particular 6-mer and a single chosen consensus site, we determined all instances of the motif, allowing for a single possible mismatch at the chosen site and matching the motif at all other sites. Considering all such aligned occurrences, we again create a co-occurrence matrix for each consensus site.

5.1.4 Motif conservation is stronger within the region of overrepresentation

A natural question to ask is whether location-specific motifs are subject to different evolutionary constraints depending upon the location in which they occur. We determined the amount of nucleotide conservation with location-specific 6mers across human and mouse, both inside the region of overrepresentation as well as within the set of intergenic sequences. Focusing on a single consensus site of a given regulatory motif, we searched genome-wide to determine the fraction of motif occurrences containing conserved nucleotides at the given site across species. We found that the majority of these consensus sites were more highly conserved within the region of overrepresentation than within the intergenic sequences (Figure 5.3). This observation agrees with those of Vardhanabhuti et al [128], and suggests that location-specific regulatory motifs are under stronger evolutionary constraints within the region at which they preferentially occur.

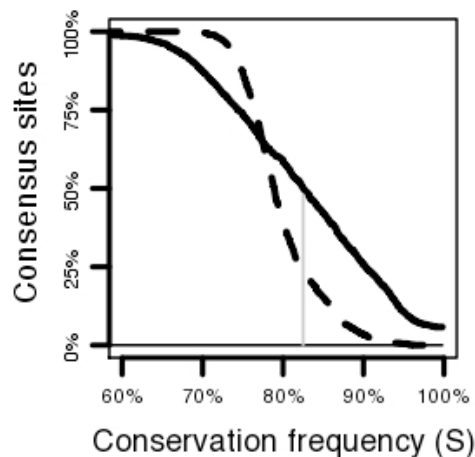


FIGURE 5.3: Conservation frequencies of location-specific regulatory motifs according to location across mouse and human. The x-axis denotes the frequency of conservation at a given consensus site of an individual regulatory motif, while the y-axis gives the cumulative number of consensus sites at or above the given frequency of conservation. The solid plot shows the amount of nucleotide conservation within the region of overrepresentation for each motif, while the dashed plot shows the amount of conservation within motif occurrences in the intergenic sequences. Note that most consensus sites tend to be more conserved within the location of overrepresentation than outside this region. At the half-way point (vertical gray line), half of all consensus sites had 83% or more conservation within the region of overrepresentation, while only 22% of the consensus sites were conserved at the same threshold within the intergenic sequences.

5.2 Modeling lineage-specific regulatory motif modifications

5.2.1 *Determining motif modifications according to nucleotide co-occurrence asymmetries*

Using the nucleotide co-occurrence matrices constructed as described previously in Section 5.1.2, we search for lineage-specific modifications within each motif by testing for differences in nucleotide preferences at each site. Given a single consensus site, we fixed one nucleotide per species at the chosen site. We then scanned for cross-species differences by comparing this frequency of nucleotide co-occurrence to that obtained after switching nucleotides across species. Large differences in these two co-occurrence frequencies indicate that each nucleotide is found frequently in exactly one

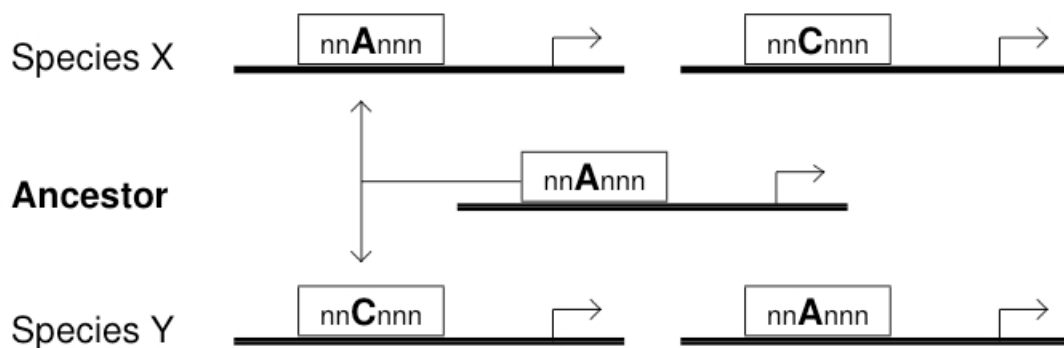


FIGURE 5.4: Regulatory motif modification scheme. Evolutionary modifications within regulatory consensus sequences involve systematic genome-wide substitutions occurring in a lineage-specific manner. This is illustrated above for a regulatory motif whose ancestral form contains an ‘A’ nucleotide at the third consensus site (center). This consensus site is systematically converted from the ancestral ‘A’ nucleotide to the newly preferred ‘C’ nucleotide in species Y (bottom left). Species X maintains the ancestral form of this regulatory motif (top left), and therefore the frequency of co-occurrence shown on the left (with A and C in species X and Y, respectively) is observed frequently across the genome. Such lineage-specific changes produce an asymmetry in nucleotide co-occurrences after switching nucleotides across species. In this case, the co-occurrence frequency shown on the right (with nucleotides C and A in species X and Y, respectively) is far less prevalent than the co-occurrences shown on the left. Such asymmetries can therefore be used to predict lineage-specific modifications within preferred regulatory sequences.

species, but is far less preferred in the other species (Figure 5.4). Such asymmetries therefore suggest global evolutionary modifications in the ancestral sequence element occurring along one of the chosen lineages.

Below we outline the statistical framework by which we determine evolutionary modifications in regulatory motifs. We used two different statistical models to assess the prevalence of lineage-specific motif modifications. The first model, denoted as the ‘intergenic background model’, uses co-occurrence frequencies within the intergenic sequences to estimate background nucleotide co-occurrence frequencies. Evolutionary modifications in the regulatory consensus motifs are then determined by comparing motif occurrences within the region of overrepresentation to these background occurrences. The second model, denoted as the ‘binomial distribution model’,

specifically considers motif occurrences only within the region of overrepresentation without the use of background occurrences in the intergenic sequences. This second approach is based upon the binomial distribution with a null hypothesis that assumes an equal probability of substitution to and from each nucleotide across lineages. We describe each of these two models below.

5.2.2 Intergenic background model

Consider two species, species X and species Y, and a single consensus site within a given motif. We search for non-identical nucleotides i and j where species X prefers nucleotide i at the given consensus site while species Y prefers nucleotide j at the same site. We set a random variable (T_{ij}) that represents the number of times nucleotide i co-occurs at this consensus site in species X along with nucleotide j at the same site in species Y. We then consider motif co-occurrences targeting orthologous genes within the region of overrepresentation across species. The observed value (t_{ij}) of T_{ij} then represents the number of co-occurrences for nucleotides i and j at the given site found genome-wide across species, as illustrated in Figure 5.2.

In order to determine rapid evolutionary modifications, we compare the value of t_{ij} to the number of nucleotide co-occurrences after switching i and j across the two species (i.e., the value t_{ji}). When no cross-species differences exist, we would expect that $t_{ij} \approx t_{ji}$, as no species-specific biases exist between the two nucleotides. In contrast, for cases in which species X has a strong preference for nucleotide i at the given site while species Y prefers nucleotide j at the same site, we would expect a significant asymmetry between these two values, giving $t_{ij} \gg t_{ji}$. In such cases, the difference $t_{ij} - t_{ji}$ will take on large positive values.

As co-occurrence data is taken across the genome and therefore provides a large sample size, we assume a normal approximation for $T_{ij} - T_{ji}$. The significance of functional asymmetry within the region of overrepresentation can then be assessed

using a Z-score. This Z-score represents the number of standard deviations by which the observed asymmetry ($t_{ij} - t_{ji}$) deviates from its expected value and is given by

$$Z = \frac{[t_{ij} - t_{ji}] - \mathbb{E}[T_{ij} - T_{ji}]}{\sqrt{\text{Var}[T_{ij} - T_{ji}]}} \quad (5.1)$$

In our model, we estimate the expected value and variance of $T_{ij} - T_{ji}$ according to a background frequency of nucleotide co-occurrence within the intergenic sequences. We determine the number of these background motif occurrence (b_{ij}) with nucleotides i and j at the given site within species X and Y, respectively. We set the total number of co-occurrences within the intergenic sequences to be N_b and the total number of co-occurrences within the region of overrepresentation to be N_t . (I.e., $N_b = \sum_{xy} b_{xy}$ and $N_t = \sum_{xy} t_{xy}$.) We show in Appendix C that the Z-score in Eq 5.1 is approximated by

$$Z = \frac{t_{ij} - t_{ji}}{\sqrt{(N_t/N_b)(b_{ij} + b_{ji})}} \quad (5.2)$$

Note that, for non-identical nucleotides i and j , the values for b_{ij} and b_{ji} reflect the background frequency of substitution at the chosen site, as co-occurrences of nucleotides i and j represent instances of non-conservation. Similarly, t_{ij} reflects a similar frequency of substitution within the area of overrepresentation. Thus, our Z-score effectively quantifies the amount of modification within the region of overrepresentation compared to that expected according to a background (intergenic) frequency of substitution. Extreme Z-scores therefore indicate rapid evolutionary changes within functional motif occurrences that are not adequately explained by the background rate of substitution.

5.2.3 The binomial distribution model

Our second approach considers only motif occurrences within the region of overrepresentation without the use of background nucleotide co-occurrence frequencies. This

approach is based upon a null hypothesis that does not require a direct comparison between motif occurrences within the region of overrepresentation to those outside this region.

For each consensus site, we fix a pair of nucleotides i and j co-occurring in species X and Y, respectively. Assuming no cross-species differences in preferred consensus sequence, the number of these co-occurrences t_{ij} should approximately equal the number of co-occurrences after switching nucleotides across species (t_{ji}). Thus, T_{ij} should follow a binomial distribution with probability 1/2 and number of trials $\hat{n} = t_{ij} + t_{ji}$. For a two-tailed test of the null hypothesis that species X and Y do not differ in nucleotide preferences, the binomial approach allows us to calculate a p-value (p) corresponding to an asymmetry between t_{ij} and t_{ji} . Equivalently, this p-value measures the deviation of t_{ij} from its expected value $(t_{ij} + t_{ji})/2$. These p-values are then converted to a ‘significance term’ P , where $P = -\log p$. Note that large values of P indicate large deviations of t_{ij} from its expected value.

5.3 Evolutionary modifications within location-specific motifs within vertebrates

5.3.1 *Genome-wide biases in nucleotide preferences are stronger within the region of overrepresentation*

We looked to assess the amount of nucleotide biases among *cis*-regulatory elements according to location in which they occur. We conducted an initial scan for cross-species differences within regulatory motifs both within and outside the region of overrepresentation between human and mouse. Differences in nucleotide preferences were quantified using a cross-species ‘asymmetry score’ Q , which we define to be

$$Q = \sum_{i \neq j} |p_{ij} - p_{ji}| \quad (5.3)$$

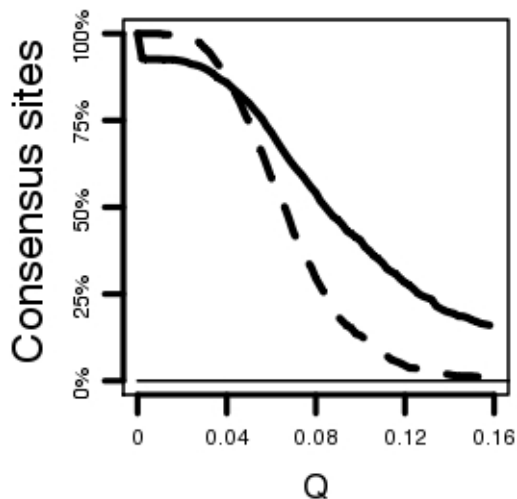


FIGURE 5.5: Asymmetry scores Q for regulatory element consensus sites within (solid plot) and outside (dashed plot) the region of overrepresentation. Q -values represent the total asymmetry of nucleotide preferences within a given consensus site, defined as $Q = \sum_{i \neq j} |p_{ij} - p_{ji}|$, where $p_{ij} = t_{ij}/N_t$ for occurrences within the region of overrepresentation and $p_{ij} = b_{ij}/N_b$ within the intergenic sequences. Note that many consensus sites show greater amounts of nucleotide biases within the region of overrepresentation compared to those outside this region.

where p_{ij} represents the frequency of nucleotides i and j across species X and Y, respectively. Namely, $p_{ij} = t_{ij}/N_t$ for motif occurrences within the region of overrepresentation, while $p_{ij} = b_{ij}/N_b$ for occurrences within the intergenic sequences.

We find that many consensus sites exhibited greater amounts of cross-species differences within the region of overrepresentation compared to occurrences within the intergenic sequences (Figure 5.5). For instance, 16% of all consensus sites had a significant amount of asymmetry at a Q -value threshold of $Q > 0.16$ within the region of overrepresentation, while less than 1% exhibited asymmetries above this threshold within the intergenic sequences.

	Human		Mouse		Genome assembly
Total	19,220 promoter sequences		18,275 promoter sequences		hg18, NCBI Build 36.1 mm9, Build 37
Species	# Orthologs	Div Time	# Orthologs	Div Time	
Human	—	—	11,270	94 mya	hg18, NCBI Build 36.1
Mouse	11,341	94 mya	—	—	mm8, Build 36
Chimp	14,764	6 mya	9,005	94 mya	panTro2, Build 2 Version 1
Macaque	14,782	30 mya	9,311	94 mya	rheMac2, assembly v.1.0
Rat	9,937	94 mya	14,277	26 mya	rn4, version 3.4
Cow	11,683	98 mya	8,425	98 mya	bosTau3, Baylor release Btau3.1
Dog	10,322	98 mya	7,789	98 mya	canFam2, assembly v2.0
Horse	9,808	98 mya	7,175	98 mya	equCab1, UCSC version equCab1
Opossum	5,377	160 mya	5,022	160 mya	monDom4
Platypus	2,395	163 mya	1,621	163 mya	ornAna1, v5.0.1
Lizard	1,589	275 mya	1,459	275 mya	anoCar1, AnoCar(1.0)
Frog	797	390 mya	925	390 mya	xenTro2, version 4.1

Table 5.2: Data sets used for cross-species comparisons. All species comparisons were conducted relative to either human or mouse. We show the number of orthologous sequences for each pair-wise comparison, as well as the amount of divergence time (in millions of years) between lineages, as estimated in [47].

5.3.2 Pair-wise species comparisons conducted within vertebrates

We scanned for patterns of evolutionary modifications within the comprehensive list of predicted 6-mer motifs exhibiting location-specific overrepresentation. Pair-wise species comparisons were conducted across a large array of twelve vertebrate species, each relative to human or mouse. Orthologous promoter sequences were determined using the genome-wide multiz28way (hg18) and multiz30way (mm9) alignments [85]. We subsequently aligned by dynamic programming to estimate the location of each orthologous TSS. Stringent quality controls were applied in order to filter low-confidence orthologs from our data set. Ortholog-pairs were excluded from the analysis if more than 25% of the pair-wise alignment columns within the (-10,+10) window contained gaps, or if less than 70% of these aligned columns contained matching nucleotides. The number of resulting ortholog-pairs for each comparison is shown in Table 5.2.

To assess the prevalence of evolutionary modifications, we applied both the intergenic background model and the binomial distribution model to each data set. Z-

and P-values were determined at each consensus site for all location-specific 6-mers and nucleotide-pair combinations i and j . In addition to single nucleotide consensus sites, we also considered doubly degenerate sites. For example, we tested for cases in which one species prefers either a ‘C’ or a ‘G’ nucleotide ($S=[C,G]$), while a second species prefers either an ‘A’ or a ‘T’ ($W=[A,T]$).

Although the vast majority of motifs produced near or over 100 non-conserved co-occurrences across species, some motifs overrepresented at very precise locations (e.g., within 1-6 nucleotide sites) produced too few motif co-occurrences to accurately characterize cross-species differences. Thus, for each species comparison, we excluded location-specific motifs producing less than 15 non-conserved co-occurrences.

5.3.3 Many cis-regulatory elements have undergone rapid evolutionary changes within mammals

Scans for evolutionary modifications within the list of predicted 6-mer motifs showed that a very significant fraction of location-specific regulatory motifs have been subject to rapid evolutionary changes in consensus sequence. Applying the intergenic background model to the co-occurrence data, we found that close to a third of all location-specific regulatory elements have been modified on either the human or mouse lineages following species divergence ($Z > 5$) (Table 5.3). Comparisons between other species with divergence times near that of human and mouse produced similar numbers of modified regulatory elements, with predicted modifications within approximately 16-35% of all location-specific motifs during the majority of the species comparisons.

The motif predictions generally appeared to be robust with regards to the RefSeq annotations used. We note that two human-mouse comparisons were conducted, one using human RefSeq promoters and another using mouse RefSeq promoters. There were 140 6-mers predicted to exhibit location-specific overrepresentation in both hu-

Motif and Consensus site predictions						
Species		Divergence time	Total Motifs	Z-score > 5		
				Motifs	Sites	
Mouse	Rat	26 mya	187	4 (2%)	4 (1%)	
	Human	94 mya	188	62 (33%)	79 (7%)	
	Chimp	94 mya	188	49 (26%)	61 (5%)	
	Macaque	94 mya	190	57 (30%)	77 (7%)	
	Cow	98 mya	188	62 (33%)	79 (7%)	
	Horse	98 mya	190	58 (31%)	78 (7%)	
	Dog	98 mya	183	88 (48%)	155 (14%)	
	Opossum	160 mya	179	99 (55%)	181 (17%)	
Human	Chimp	6 mya	143	9 (6%)	9 (1%)	
	Macaque	30 mya	168	11 (6%)	11 (1%)	
	Mouse	94 mya	177	52 (29%)	69 (7%)	
	Rat	94 mya	174	61 (35%)	82 (8%)	
	Cow	98 mya	174	31 (18%)	32 (3%)	
	Horse	98 mya	173	28 (16%)	29 (3%)	
	Dog	98 mya	174	41 (24%)	52 (5%)	
	Opossum	160 mya	168	78 (46%)	120 (12%)	

Table 5.3: Fractions of regulatory motifs and consensus sites exhibiting evolutionary modifications within mammals. For each pair-wise species comparison, we show the number and fraction of location-specific motifs exhibiting rapid evolutionary modifications ($Z > 5$) as well as the fraction of consensus sites with nucleotide preferences differing across species. The amount of divergence time between species as well as the number of location-specific motifs considered during each species comparison (i.e., the number of motifs with at least 15 non-conserved co-occurrences across species; see Section 5.3.2) are also shown.

man and mouse. Among these common 6-mers, thirty-four (24%) were predicted to be modified using the human RefSeq data, and thirty-seven (26%) were predicted to be modified using the mouse RefSeq data. Twenty-two of these modified motif predictions overlapped between the two species comparisons. This overlap is highly significant, comprising 65% of the motif predictions from the human RefSeq data. Many of the remaining (non-overlapping) motifs also produced high Z-scores, albeit under the stringent threshold of $Z > 5$. Of these overlapping predictions, the majority (59%) comprised reciprocal consensus site modifications, for which the same consensus site and nucleotide-pair were predicted using both the human and the mouse RefSeq promoter data.

5.3.4 *Correlation between divergence time and motif modification*

We noted a strong correlation between divergence time and the prevalence of regulatory motif modification (Figure 5.6). Only 6% of the motifs exhibited significant differences within primates, which share common ancestry within 30 million years ago [47]. Similarly, very few motifs were found to differ between mouse and rat (divergence time ~ 26 million years [47]), with only 2% of all location-specific motifs exhibiting cross-species differences within rodents. In contrast, about half of all location-specific motifs exhibited differences between eutherians and the more distantly related opossum lineage (divergence time ~ 160 million years [47]). This correlation was strongly linear ($p < 1e-5$), suggesting a relatively constant rate of modification during mammalian evolution.

5.3.5 *Simulation analyses and the effect of multiple hypothesis testing*

In our scan for motif modifications, we tested each of the six consensus sites and nucleotide pair i and j across all location-specific motifs. It is therefore natural to inquire about the effect of multiple hypothesis testing and the expected number of false-positives within our list of predictions. In order to assess the effect of multiple hypothesis testing as well as the effects of random fluctuations within the data, we conducted simulation analyses by randomizing nucleotide co-occurrences for each pair of nucleotides across species. Specifically, for each motif co-occurrence producing non-identical nucleotides i and j in species X and Y, respectively, we randomly chose to switch nucleotide j to species X and nucleotide i to species Y with probability $1/2$. Observed co-occurrences of j and i in X and Y, respectively, were randomly switched in a similar fashion. Note that this preserves the total number of co-occurrences of i and j across the two species, albeit enforcing a uniform probability of substitution towards each nucleotide in both species. This amounts to simulating data sets according to the null hypothesis that i and j randomly co-occur across

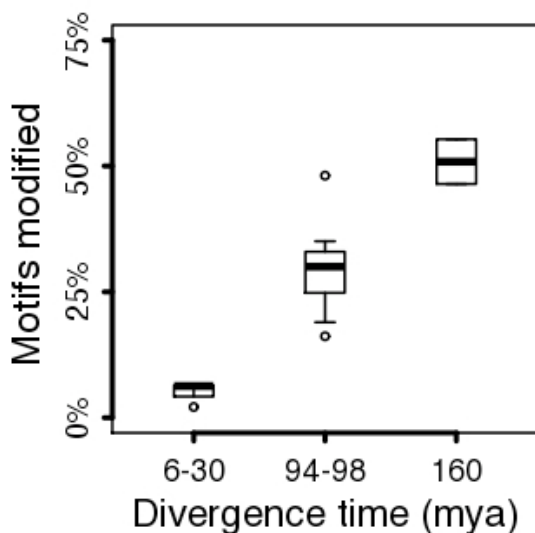


FIGURE 5.6: Prevalence of rapid evolutionary modifications within location-specific regulatory motifs according to divergence time. y-axis values represent the fraction of regulatory motifs exhibiting rapid evolutionary changes in consensus sequence following species divergence (Z -score > 5). Species comparisons were conducted in a pair-wise fashion, each comparison producing a single set of modified motifs. The sets of modified motifs are separated according to divergence time between the corresponding pair of species (x-axis) [47]. Each barplot shows the median fraction of modified motifs (center line), the first and third quartile (bar extremes), and the most extreme comparisons within 80% of the inter-quartile range (standard bar), while circles represents single outlier points.

species without lineage-specific differences in nucleotide preference.

Our analysis was then conducted using the randomized data set in an identical fashion to the real data analysis. We scanned for spurious instances of motif modification at each site of all location-specific motifs and nucleotide pairs i and j , repeating the process three times for each species comparison and re-randomizing the data during each of the three analyses.

Very few motifs were predicted using the randomized co-occurrence data, suggesting few false-positives within our set of predictions. Across all three analyses and species comparisons, our model produced a total of 2 motif predictions using the randomized co-occurrence data. This gives an estimated false-positive rate of approximately $3e-8$, similar to that expected given the assumption that random

asymmetries follow a Gaussian distribution as in our null hypothesis (about $3e-7$). Thus, we can be confident that our highly conservative Z-score threshold ($Z > 5$) was sufficiently stringent to account for multiple hypothesis testing.

5.3.6 The high prevalence of motif modification is supported by the binomial distribution model

Here, we are less interested in providing a definitive statistical model to predict motif modifications than collecting evidence that such modifications are in fact prevalent among *cis*-regulatory elements. Thus, we tested whether the alternative binomial distribution model (Section 5.2.3) would support our previous results. We performed a second scan for regulatory motif modifications according to this model, which considers only motif occurrences within the region of overrepresentation, without the use of the background occurrences. A P-value threshold was set to that producing less than a 5% false positive rate according to our randomized simulation analysis. This value was empirically found to be $P > 6.5$.

Results from this alternative statistical model confirmed the high prevalence of evolutionary motif modification. For instance, approximately 27% of all location-specific motifs were predicted to exhibit evolutionary modifications since the divergence of human and mouse. Many of these predicted motifs overlapped with those predicted during the previous analysis, with fifteen of the top 20 motif predictions also producing significant Z-scores in the previous scan for modifications. Overall, the majority of comparisons between species with similar divergence times as human and mouse predicted between 13-33% of all location-specific motifs to exhibit modifications, with only a few exceptions (Table 5.4). These results confirm the high prevalence of evolutionary modifications within our set of predicted motifs, and suggest that regulatory element modifications can be detected using statistical methods based upon different underlying assumptions.

Binomial distribution model predictions				
Species		Divergence time	Total Motifs	Predictions P > 6.5
Mouse	Rat	26 mya	187	13 (7%)
	Human	94 mya	188	57 (30%)
	Chimp	94 mya	188	24 (13%)
	Macaque	94 mya	190	30 (16%)
	Cow	98 mya	188	31 (16%)
	Horse	98 mya	190	33 (17%)
	Dog	98 mya	183	97 (53%)
	Opossum	160 mya	179	61 (34%)
Human	Chimp	6 mya	143	2 (1%)
	Macaque	30 mya	168	10 (6%)
	Mouse	94 mya	177	44 (25%)
	Rat	94 mya	174	57 (33%)
	Cow	98 mya	174	13 (7%)
	Horse	98 mya	173	14 (8%)
	Dog	98 mya	174	60 (34%)
	Opossum	160 mya	168	56 (33%)

Table 5.4: Numbers and fraction of location-specific motifs predicted using the binomial distribution model. A prediction threshold of $P > 6.5$ ($P = -\log p$ for p-value p) according to a two-tailed binomial test was used during the analysis.

5.3.7 Both site degeneracy and preferred consensus nucleotides change over the course of evolution

Inspection of the results showed that, in many cases, motifs with evolutionary modifications exhibited differences in degeneracy. For instance, in the case of the TATA-box motif (TATAAA), the degenerate ‘G’ nucleotide was far more prevalent at the first site in non-rodents than in mouse. None of the non-conserved TATA-box occurrences contained a ‘G’ at this site in mouse, although it was commonly found among all non-rodent lineages. This suggests a branch-specific change along this lineage following the eutherian radiation.

In addition to changes in the amount of degeneracy, there were also numerous cases in which the most common nucleotide sequence differed across lineages. Between human and mouse, about 16-19% of all location-specific motifs differed in their most common consensus sequences. Similar numbers of motifs (approximately

	Mouse	Human	Z-score
1	cggg M g	cggg G g	9.02
2	cc G cct	cc C cct	8.79
3	gggc G t	gggc C t	8.11
4	ctgcc Y	ctgcc G	7.57
5	tggg C g	tggg G g	7.38
6	g C ggag	g G ggag	6.98
7	c T gctg	c G gctg	6.37
8	tgc G gc	tgc C gc	5.49
9	g Y ccgc	g G ccgc	5.33
10	cgga S a	cgga A a	5.25

Table 5.5: Ten regulatory motifs with consensus sequence differences between human and mouse.

13-22%) showed cross-species differences across other eutherians. In several cases, modifications in the most frequent nucleotide consensus sequence were found to occur at rapid rates, producing significant Z-scores. Ten such motifs differing between mouse and human are given in Table 5.5. We note that, although many studies have analyzed cross-species conservation between eutherian lineages such as human and mouse [146, 128], modification in preferred regulatory consensus sequences appears to be relatively common along these lineages.

5.3.8 *The GC box regulatory motif has been modified along the eutherian branch*

Many motifs exhibited differences between eutherians and non-eutherian vertebrates. One notable example is the SP1 binding site, commonly referred to as the GC box. We find that the previously studied form of this element (gggCgg) in fact represents an altered version of its ancestral sequence, consistently found among non-eutherians as the gggAgg consensus sequence. As the term ‘GC box’ was derived from its well-studied consensus sequence, we refer to the predicted ancestral sequence as the ‘GA’ box. There is a striking pattern of GC/GA box co-occurrences between eutherians and non-eutherians, respectively, with the ancestral form commonly appearing in lineages ranging from opossum to *Xenopus* (Table 5.6). Interestingly, this pattern was found upon both strands of this regulatory element.

Evolution of the GC box motif						
Human	Mouse			Mouse		
		gggCgg	gggAgg		ccGccc	ccTccc
	gggCgg	49%	4%	ccGccc	45%	4%
	gggAgg	5%	20%	ccTccc	5%	23%
Opossum	Mouse			Mouse		
		gggCgg	gggAgg		ccGccc	ccTccc
	gggCgg	28%	5%	ccGccc	28%	5%
	gggAgg	17%	20%	ccTccc	15%	23%
Platypus	Mouse			Mouse		
		gggCgg	gggAgg		ccGccc	ccTccc
	gggCgg	16%	2%	ccGccc	12%	3%
	gggAgg	11%	27%	ccTccc	13%	29%
Lizard	Mouse			Mouse		
		gggCgg	gggAgg		ccGccc	ccTccc
	gggCgg	32%	7%	ccGccc	23%	3%
	gggAgg	18%	15%	ccTccc	27%	13%
Frog	Mouse			Mouse		
		gggCgg	gggAgg		ccGccc	ccTccc
	gggCgg	21%	3%	ccGccc	26%	6%
	gggAgg	21%	24%	ccTccc	22%	20%

Table 5.6: Evolutionary changes within the GC box element. Comparisons were made between GC box occurrences in its well-studied form (gggCgg) compared to its ancestral sequence as predicted during our analyses (gggAgg, denoted here as the ‘GA’ box). The fraction of co-occurrences targeting orthologous genes for each combination of these two versions is given for mouse relative to five other vertebrate species. We note a strong trend for the GC box to occur frequently within mouse along with the GA box within non-eutherians, as shown in the lower-left entry of each matrix. In contrast, the frequency of co-occurrence is much lower when switching these motifs across species (top-right matrix entries). This pattern was found for both the forward strand (left) and the reverse strand (right) of the motif. In contrast, no significant difference in nucleotide preference is found strictly within eutherians, illustrated above by the mouse-human comparisons. This suggests a lineage-specific modification of the ancestral GA box regulatory element to the GC box form along the eutherian branch following the split with marsupials.

The consistency of the preferred GA box motif within non-eutherians indicates that this regulatory element was modified following the split with opossum but prior to the divergence of the various eutherians lineages. Separate analyses conducted upon zebrafish showed significant amounts of locational overrepresentation of the ancestral form but no locational specificity of the GC box form, suggesting that the common eutherian version of this regulatory element is largely non-functional along the zebrafish lineage.

As this regulatory motif is highly prevalent among nearly all vertebrates, the conversion of the ancestral form to its common eutherian version represents sequence modifications across hundreds of functional sites genome-wide. Over 9% of all orthologous target genes contained a GC/GA box co-occurrence between mouse and opossum, respectively, while the reverse co-occurrence was only about half as common. As it has been estimated that opossum and eutherians share approximately 15,000 orthologous genes [83], we estimate a genome-wide difference of ~ 600 more genes containing a GC/GA box co-occurrence in eutherians and non-eutherians, respectively, than vice versa. The rate of modification for this element was particularly high relative to the background rate, producing a Z-score of $Z = 8.5$ during the mouse-opossum comparisons.

5.4 Discussion

Our results represent a novel form of molecular co-evolution, as most studies regarding macro-molecule co-evolution focus only upon compensatory changes within specific genes or pairs of genes. For example, although molecular co-evolution has been documented between interacting protein-pairs [40] as well as within RNA complementary base-paired regions [96], such changes naturally occur either within or between individual genes. In contrast, the regulatory motif modifications observed here involve a massive number of changes occurring across the entire genome, poten-

tially induced by modification within a single *trans*-factor.

Because of the ‘one-to-many’ relationship between a single *trans*-factor protein and the hundreds of functional sites to which it binds, such changes have previously been assumed to occur only rarely [31, 44, 60, 61, 72, 128, 146], and therefore have thus far gone unrecognized. These results suggest that the genomic landscape is highly adaptable, as modifications in preferred consensus sequences reflect a very high number of substitutions accumulating independently across the genome. In some cases, such as the modification of the GC/GA box regulatory element described above, this can represent multiple hundreds of substitutions occurring in a relatively short period of time along a particular lineage. The ability of the genome to accommodate such coordinated modifications is surprising and perhaps counter-intuitive, yet our results suggest that such modifications have occurred relatively frequently over the course of vertebrate evolution.

5.4.1 *Use of location-specific regulatory motifs*

Locational specificity is a convenient characteristic that can be utilized to study regulatory element evolution, as our results, as well as those from previous studies [119, 128], suggest that motifs often serve specific functional roles when they occur within the region of overrepresentation. This therefore allows us to focus upon functional occurrences that are likely to play a role in gene regulation, as determined according to their location within the promoter. Furthermore, this feature can be used to distinguish these functional occurrences from the remaining ‘background’ occurrences exhibiting weaker evolutionary constraints.

It is likely that the same modifications are also prevalent within regulatory motifs without locational specificity, such as those occurring within distal enhancers. However, distinguishing between functional and putatively non-functional occurrences of these regulatory motifs is less straight-forward, as many *trans*-factor proteins bind in

concert to interacting proteins or according to topological characteristics of the DNA [45, 134, 144]. In addition, many *cis*-regulatory elements functioning within distal enhancers may only target a small handful of genes, in contrast to the frequently occurring functional occurrences of location-specific motifs. Due to this characteristic, a genome-wide approach is less applicable, and it is difficult to predict evolutionary modifications using a small number of functional element occurrences. Although the prevalence of evolutionary modifications within motifs occurring in distal enhancers has not yet been assessed, there is no reason to assume *a priori* that the high prevalence of evolutionary modification is unique to motifs functioning within the proximal promoter region.

5.4.2 Mechanisms of *cis*-regulatory element modification

We can imagine that, through various mechanisms, the observed evolutionary modifications in consensus sequence would occur. One potential mechanism may simply involve changes within the binding domain of the corresponding *trans*-factor. Such changes would likely induce nucleotide-specific substitutions genome-wide, preserving the *trans*-factor's ability to bind near the same set of target genes. In such cases, we note a dichotomy between functional preservation and sequence conservation, as it is likely that sites converted to the modified sequence element would continue to recruit the *trans*-factor, preserving the original function. In contrast, sites conserved in sequence would likely lose the ability to bind the altered protein factor, potentially affecting the expression patterns of the target gene.

At the present time, it is difficult to directly map the co-evolution of *cis*-regulatory elements and their binding *trans*-factors, and it is plausible that some of the motif modifications observed may be caused by other mechanisms. For example, paralogous or alternatively spliced proteins may share similar DNA binding sequences, yet distinct binding modules. Multiple versions of such *trans*-factors may be lost, gained,

or modified during evolution; such mechanisms are thought to have a significant impact upon evolutionary changes in morphology [52, 89]. In addition, protein-protein and protein-DNA interactions can often alter the conformation of a given binding protein [26, 66], and thus interacting factors that do not directly bind to the DNA may also cause changes in DNA binding sequences. In some cases, differences in effective population size between lineages may also produce differences among species in commonly occurring motifs, particularly in the presence of strong mutational biases towards a particular nucleotide [8]. Although these latter changes may not necessarily reflect lineage-specific differences in relative fitness, they are still likely to have direct consequences upon the expression of the affected target genes. These mechanisms are all likely to have a substantial impact upon species morphology, and embody a previously unexplored aspect of genome evolution.

6

Conclusions

6.1 Summary

This work utilizes locational specificity as a means by which to predict *cis*-regulatory elements, spatial relationships between pairs of motifs, as well as the evolution of *cis*-regulatory elements. We show that locational specificity within or near the promoter is common among *cis*-regulatory elements, and it is a useful characteristic that can be used to study regulatory motif evolution on a genome-wide scale. We generalize this model to study spatial characteristics between pairs of regulatory motifs, and we show that this provides a powerful means by which to predict functional relationships between *cis*-regulatory elements.

Although a few studies have utilized location-specific overrepresentation to predict transcription factor binding sites *de novo* [38, 119, 128], the prevalence of this characteristic among *cis*-regulatory elements has previously been unclear. Some studies appear to have largely under-estimated the prevalence of locational specificity, with FitzGerald et al [38] detecting a total of nine motifs exhibiting location-specific overrepresentation in humans. In contrast, we predict significant amounts of locational specificity in 48 non-redundant putative regulatory elements. We have shown

that our methodology allows for a significant increase of sensitivity by considering instances of location-specific overrepresentation occurring across broad or narrow regions within the proximal promoter, ranging from those occurring at a single site to those occurring across a hundred or more nucleotide sites. In contrast, most low-resolution approaches have previously been unable to detect locational specificity in even the most well-known regulatory motifs, such as the TATA-box [128] and the Inr sequence [38, 128]. In other cases, previous scans for location-specific regulatory motifs have neglected the effects of dinucleotide fluctuations near the TSS, and thus greatly over-estimate the number of GC-rich regulatory motifs [119]. Our methodology effectively accounts for fluctuations in dinucleotide content across the promoter, improving specificity and filtering spurious GC-rich motif predictions from the results.

We extend our model to analyze distance preferences between pairs of motifs, predicting putative *cis*-regulatory elements that bind interacting protein factors. It is known that transcription factors generally function in concert with other interacting factors [4, 10, 33, 144], yet few large-scale analyses regarding functional binding site relationships have been conducted. Several studies have shown that relationships do exist between pairs of regulatory motifs [5, 59, 70, 75, 95, 112, 118], although such studies have generally been limited to motifs whose sequence composition has previously been known. Although previous approaches have been effective at demonstrating that inter-relationships exist between regulatory elements, very few computational methods have been available to predict motif-pair relationships *de novo*.

Here, we use our model predict motif-pairs binding interacting proteins according to spatial preferences between pairs of sequence elements. Our methodology allows us to conduct such analyses without any prior knowledge regarding the sequence composition of the regulatory elements involved. Using a genome-wide approach, we

show that many regulatory motif-pairs exhibit inter-motif distance preferences.

Previously, it has been shown that some protein-protein interactions occur in a periodic nature, with multiple preferred separation distances characteristically phased according to the turn of the DNA double-helix or around the histone complex [53, 69, 135, 144]. Our results support the conclusions of these studies, and we show that motif-pairs binding interacting factors have multiple distances at which they preferentially occur. In our application of the model, we predict functional relationships between regulatory motifs using this ‘interaction phasing’ characteristic. We observe that many motif-pairs exhibit periodic spatial preferences, with the distance between preferred inter-motif distances exhibiting consistent phasing intervals. We find that the phasing intervals often correspond approximately to the number of nucleotides in a single turn of the DNA double-helix. It is likely that this reflects a requirement for the transcription factor-pairs to bind in a consistent orientation relative to the turn of the DNA molecule. However, we also note that many phasing intervals deviate slightly from the exact number of nucleotides in a turn of the double helix (e.g., 8-10 bp versus 10.5 bp). It is likely that this observation reflects a distortion of the DNA upon protein binding, which has often been observed during structural analyses of protein-DNA complexes [54, 92, 100]. We observe that several regulatory motifs exhibit consistent phasing intervals across a wide range of partner motifs, and we utilize this characteristic to predict putative functional relationships between binding site motifs. Our analysis resulted in a total of 112 non-redundant motif-pair predictions. We show that many of the predictions correspond to motifs binding proteins with known interactions, while the remaining predictions comprise putatively novel functional motif-pair relationships.

In this work, we assess the prevalence of evolutionary modifications in location-specific motifs across a wide array of vertebrate species. Virtually all cross-species comparison studies have previously assumed that preferred protein binding sequences

change little over time, and that such modifications occur at a negligible rate during the course of evolution [31, 44, 60, 72, 128, 141, 146]. In contrast to this assumption, we find that a surprisingly high number of the predicted regulatory motifs have been subject to significant modifications over the course of vertebrate evolution, even among eutherians with divergence times similar to that of human and mouse. Such modifications occur in a lineage-specific manner, often at rapid rates of change across the genome. In some cases, the number of nucleotide-specific substitutions, such as those for the SP1 binding site, can reach into the hundreds genome-wide. This represents a previously unrecognized mechanism driving genome evolution, and has broad implications regarding evolutionary changes in phenotype.

6.2 Future directions

In future work, we plan to expand our study of location-specific regulatory element evolution. Our previous work represents only an initial survey of the prevalence of motif modification, yet several questions remain regarding the nature of regulatory element evolution. Such questions include:

1. What is the relationship between regulatory motif modification and changes in gene expression?
2. What is the nature of *cis*- and *trans*-element co-evolution? Are significant modifications in DNA binding motifs due primarily to large modifications in protein factors, or are they caused by subtle changes in protein sequence or structure?
3. What is the role of *trans*-factor paralogs or alternative splicing in DNA binding sequence modifications? Are binding motifs more likely to become modified if several similar, yet distinct, versions of a protein are present within a given lineage?

4. Can we predict, either through sequence phylogeny or structural methods, which evolutionary changes within a transcription factor produce cross-species differences in preferred binding sequences?

Effects of regulatory motif modification on expression patterns. A natural question to ask is simply why genome-wide conversions of regulatory motifs occur, particularly to the extent at which we observe in our results. It is reasonable to hypothesize that such site conversions reflect a general preservation of the target genes bound by the given *trans*-factor, although we have yet to confirm this hypothesis empirically. One potential approach to answer this question would simply be to analyze expression patterns of orthologous genes across species, and determine the amount of similarity in expression patterns in ‘converted’ versus conserved sites. If the conversion at the non-conserved sites does, in fact, reflect a preservation of the target genes, then we would expect a conservation of gene expression patterns of the converted genes. In contrast, we would expect that conserved sites would lose the ability to recruit the *trans*-factor, and thus we may observe a change in gene expression patterns. However, further work must be done to determine whether or not this is the case.

The nature of regulatory protein/binding element co-evolution. As evolutionary modifications within regulatory elements represents a largely unexplored area of study, very little is known regarding the nature of *cis*- and *trans*-regulatory element co-evolution. For instance, one unanswered question is whether changes in protein binding affinities are primarily due to significant modifications in sequence or structure of the given *trans*-factor, or whether significant changes in binding affinities are more often due only to subtle changes in the protein’s amino acid sequence. If the former possibility is true, then we would expect that several changes are necessary to accumulate within the amino acid sequence in order to modify the protein’s binding affinity. However, in such a case, the co-evolution of the *cis*- and *trans*-regulatory el-

ements would occur in a step-by-step manner. This possibility appears unlikely given our results, as such a series of incremental changes would be likely to accumulate across the entire phylogeny. This is not what we observe; instead, motif modification appears to occur rapidly and usually along a single lineage. This may indicate that even significant changes in binding affinities are often caused by subtle changes in the corresponding *trans*-factor. However, due to the one-to-many relationship between a transcription factor and its many binding sequences, it is not clear why such subtle changes would occur, as they must then induce hundreds of *cis*-element substitutions across the genome. Thus, such questions remain, and the nature of co-evolution between *cis*- and *trans*-elements has yet to be fully understood.

Multiple versions of related proteins and their effects upon motif modification. There is a growing body of evidence that multiple versions of related proteins, such as protein paralogs and alternatively spliced proteins, have major effects upon the evolution of species morphology [52, 89]. It is unclear, however, how different forms of a protein factor affect *cis*-regulatory element evolution, and how this ultimately affects phenotype. It is possible, albeit unknown, that the ‘one-to-many’ *trans*-to-*cis*-element model is oversimplified, as similar binding domains within related proteins may bind to similar motifs. Divergence of these different protein versions can cause modifications in structure and binding affinities, although the extent to which this produces modifications in *cis*-regulatory elements, in comparison with the one-to-many model, is unknown. Such effects may well be biologically relevant, although these phenomena have not yet been studied.

Mapping trans-regulatory protein changes to changes in binding affinities. A complete understanding of the *cis*-regulatory element modifications presented here involves knowledge regarding the changes in sequence and structure of the corresponding binding proteins. Not only does this provide insights into the study of evolution, but it is also generally applicable to any area of research regarding macro-

molecule interactions. However, whether we are interested in molecular interactions or in evolutionary mechanisms *per se*, comparative genomics offers a unique approach to study the nature of protein-DNA complexes. The advantage is two-fold. First, we can apply phylogenetic sequence data to predict candidate sites potentially involved in the modification of the corresponding binding sequence. Second, we can use a structural approach, either computational or experimental, using orthologous proteins collected from different species to predict changes in potential energy within the binding complex. In our case, the use of cross-species comparisons facilitates our study into molecular interactions, since we are able to focus upon related proteins that have been computationally predicted, rather than searching for differences across an large set of randomly chosen protein structures. From an evolutionary perspective, it would be of major interest to predict which changes in the protein modify its binding sequence and which changes are non-functional. Creating mathematical models to study such changes computationally and assessing the validity of such methods would have broad impacts upon across a wide range of studies.

6.3 Concluding Remarks

The rapid increase of available genomic data has made it possible to study fundamental biological processes using DNA sequence analysis. Our approach to detect and study *cis*-regulatory elements uses DNA sequence data on a genome-wide scale, and presents a powerful means by which to utilize high-throughput sequence data. Although we do not expect that all biological questions can be answered using sequence analysis alone, it provides a step forward towards our knowledge of the various molecular mechanisms that ultimately affect organismal phenotype. *In silico* approaches in general can be used both to focus future research upon the computationally derived predictions, and also to understand the underlying biological mechanisms that operate within the organism. It is our hope that future work will demonstrate the

effectiveness of the approach presented here, and provide further motivation towards utilizing spatial preferences as a means by which to study gene regulation.

Appendix A

MLF methodology

A.1 Statistical framework of the MLF model

Below we describe the specifics of the MLF model. Our goal is to predict biologically relevant instances of location-specific overrepresentation, and to use this characteristic as a criterion for regulatory function. The method is an application of non-linear regression: given a set of observed motif occurrences within a set of promoter sequences, we collectively estimate the underlying frequency of occurrence according to location. Locational overrepresentation is modeled in a continuous fashion. As described briefly above, for a given motif w , its MLF $g_w(x)$ represents the underlying probability of occurrence according to its position x . Suppose our data set s consists of N sequences, each of length L . Thus, we have $s = \{s_1, s_2, \dots, s_N\}$ where $s_i = s_i(1)\dots s_i(L)$. We consider these sequences to be the observed outcome of an underlying biological process, and define a random variable S analogous to a single sequence in s , where $S = S(1)S(2)\dots S(L)$. We then define a random variable $U_k(j)$ to be the kmer starting at position j in S : $U_k(j) = S(j)S(j+1)\dots S(j+k-1)$. Our

model then defines the MLF $g_w(x)$ to be

$$g_w(x) = \Pr(U_{l_w}(x+t) = w) \quad (\text{A.1})$$

where x represents the position of the motif, t represents the position of the TSS, and l_w represents the length of w . Note that the position x is relative to t , and thus the location of the TSS is given by $x = 0$. The value of $g_w(x)$ represents, for any individual position x , the underlying probability of occurrence of w at this precise location. The values of this function are not normalized across the values of x , and therefore the sum of the values do not, in general, equal 1 across the promoter.

As discussed previously, MLFs are modeled as the sum of a ‘background function’ $C(x)$ and a ‘signal function’ $H(x)$. Thus, for any MLF g , we have

$$g(x) = C(x) + H(x) \quad (\text{A.2})$$

The background function $C(x)$ represents the background frequency of the motif, namely, the frequency without explicit locational specificity. This function is allowed to fluctuate according to the dinucleotide makeup of the promoter and is modeled as a polynomial (see below). In contrast, the signal function $H(x)$ incorporates possible locational overrepresentation into the model. We remember that the signal function is modeled as a single unnormalized Gaussian term times a coefficient a :

$$H(x) = a \cdot \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right] \quad (\text{A.3})$$

Here, the parameters a , μ , and σ are free parameters that are estimated according to trends observed within the promoter data; these parameters are determined using likelihood-maximization as described below in Section A.3.

A.2 Background functions

The background function $C(x)$ represents the background frequency of the motif at position x . This function is estimated using a ‘prototype background function’ ($c(x)$)

which represents the expected frequency of occurrence. This expected frequency is determined according to the dinucleotide composition at each position within the promoters. We distinguish between the background frequency $C(x)$ and the ‘expected’ frequency of occurrence $c(x)$, since most motifs are either over- or under-represented with respect to their dinucleotide makeup. The underlying expected frequency is modeled using a polynomial function:

$$c(x) = \sum_{k=0}^K h_k x^k \quad (\text{A.4})$$

This function is obtained by conducting linear regression on the set of data points $\{\langle x, R(x) \rangle\}$, where $R(x)$ represents the expected frequency of occurrence at position x . The value $R(x)$ is determined independently at each nucleotide site according to the observed dinucleotide frequencies at that particular location. The function $c(x)$ then gives the underlying probability of occurrence after fitting a polynomial to this data set. The degree of this polynomial (K) is unique to each motif, and reflects the expected amount of fluctuation in occurrence frequency according to changes in dinucleotide content across the promoters.

Formally, $R(x)$ for a motif w of length l_w , i.e., $w = w(1) \dots w(l_w)$, is given by a position-specific 1st order Markov-dependency model as described in Karlin *et al.* [57]. The expected frequency $R(x)$ of w at position x is given by

$$R(x) = \frac{\prod_{i=1}^{l_w-1} R_{w(i)w(i+1)}(x+i-1)}{\prod_{i=2}^{l_w-1} R_{w(i)}(x+i-1)} \quad (\text{A.5})$$

where $R_{w(i)w(i+1)}(x)$ gives the observed frequency of the dinucleotide $w(i)w(i+1)$ at position x , and $R_{w(i)}(x)$ represents the analogous mono-nucleotide frequency.

Since the expected frequency of many motifs differs from the actual frequency of occurrence, the background frequency $C(x)$ of a kmer w is allowed to deviate from

$c(x)$. Namely, we allow, for uniformly distributed over- and under-representation. We therefore model $C(x)$ as

$$C(x) = b + d \cdot c(x) \tag{A.6}$$

where b and d are free parameters. Thus, the background model is allowed to ‘shift’ and ‘stretch’ vertically using parameters b and d , respectively; this allows for uniformly distributed differences in the expected and observed occurrence frequencies.

A.3 Parameter estimation and statistical significance determination

As noted briefly above, parameter estimates (i.e., b , d , a , μ , σ) are obtained using likelihood-maximization. For the model assuming no locational overrepresentation (i.e., $H(x) = 0$), only the parameters b and d must be estimated. In this case, since both b and d are linear parameters within the model, the likelihood can be maximized directly using linear regression. For models incorporating locational specificity, however, the parameters must be obtained through non-linear regression analysis. This is done by optimizing the log-likelihood $L(D; \theta_g)$ of the data D given the model g , where θ_g is the parameter vector of model g . Here, the data set D is given by: $D = \{\langle x_1, z_1 \rangle, \dots, \langle x_n, z_n \rangle\}$, where z_i represents the number of motif occurrences at position x_i . As each MLF $g(x)$ represents the probability of motif occurrence at x , the log-likelihood of a single data point $L(\langle x_i, z_i \rangle; \theta_g)$ reflects the outcome of multiple Bernoulli trials with a ‘success’ being an occurrence of the motif at position x_i . Thus, the log-likelihood of this data point is given by the binomial distribution:

$$L(\langle x_i, z_i \rangle; \theta_g) \equiv z_i \cdot \log[g(x_i)] + (N - z_i) \cdot \log[1 - g(x_i)] \tag{A.7}$$

where N is the number of sequences within the data set (i.e., the number of ‘trials’). The total log-likelihood $L(D; \theta_g)$ of the data is given by the sum of the log-likelihoods across all data points $\langle x_i, z_i \rangle$. This value is maximized using an iterative method

called ‘Broyden’s method’ [14] given an initial parameter estimate θ_0 . Several initial parameter vectors are used during each MLF estimation; the final parameter estimates are taken to be those producing the highest log-likelihood. Initial parameter vectors are chosen according to outlier data points whose deviation from the background frequency may suggest overrepresentation at the specific location in which they are found.

A.4 Model selection

Model selection for any given MLF involves determining both the degree K of the prototype background function as well as the existence of a Gaussian term (i.e., either $a = 0$ or $a \neq 0$ in Eq A.3). These are determined in the same manner; namely, we use a likelihood ratio test (F-test) to compare the log-likelihoods of the data given two possible models. To determine the presence or absence of locational overrepresentation, we compare the log-likelihood derived from the (null) model where $H(x)$ is identically zero to that of the (alternative) model where $H(x)$ takes on non-zero values. Model selection involves comparing the log-likelihoods $L(D; \theta_{g_A})$ and $L(D; \theta_{g_0})$, where the MLF g_A allows for locational overrepresentation, while its nested model g_0 assumes no locational bias. The ‘scaled deviance’ $Z(\theta_{g_A}, \theta_{g_0})$, given by

$$Z(\theta_{g_A}, \theta_{g_0}) = 2 \cdot [L(D; \theta_{g_A}) - L(D; \theta_{g_0})] \quad (\text{A.8})$$

follows a χ^2 distribution with $|\theta_{g_A}| - |\theta_{g_0}|$ degrees of freedom [29].

Our final statistic is

$$F = \frac{Z(\theta_{g_A}, \theta_{g_0}) \cdot (n - |\theta_{g_A}|)}{Z(\theta_S, \theta_{g_A}) \cdot (|\theta_{g_A}| - |\theta_{g_0}|)} \quad (\text{A.9})$$

where n is the number of data points and model S is the ‘saturated model’, i.e., the model optimizing the log-likelihood at each data point without limits on the number of parameters. The value F follows the F-distribution with $|\theta_{g_A}| - |\theta_{g_0}|$ and

$n - |\theta_{g_A}|$ degrees of freedom [29]; p -values reflecting the significance of locational overrepresentation are derived using this statistic.

Determining the order K of the polynomial $c(x)$ is also determined using an F-test. This is done in an incremental fashion. Specifically, we first determine whether $c(x)$ is non-constant by considering the possibility that $K = 1$, and we compare the likelihood of this model to that in which $K = 0$. Comparisons between these two possibilities are again conducted using an F-test, where significant p -values produced from this test indicate that $c(x)$ is non-constant (i.e., $K \geq 1$), while non-significant p -values suggest that $c(x)$ is uniform across the promoter (i.e., $K = 0$). For cases in which $c(x)$ is non-constant, we then increment the value of K , comparing models where $K = 2$ to that in which $K = 1$. Again, this comparison is conducted using an F-test, where significant p -values indicate that $c(x)$ is non-linear (i.e., $K \geq 2$). This process is repeated, incrementing the degree K until the F-test no longer produces a significant p -value. The final value of K is then taken to be the last value of K that produces a significant p -value; here, we set the p -value threshold to be $p < 1e-5$.

A.5 Motif clustering procedure

For the MLF clustering analyses, 6mer motifs are clustered for redundancy according to both sequence similarity as well as the location and width of overrepresentation. Clustering is conducted by considering 6mers in rank order. At each step, an individual 6mer motif is either placed in an existing cluster or else a new cluster is created. 6mers matching at five of the six sites (i.e., containing only one mismatch, or no mismatches with a single bp offset) are clustered if their signal functions are similar according to their KL divergence [65]. The KL divergence between two signal functions is calculated by converting each function into a discrete probability

distribution $p(x)$ across each position within the promoter:

$$p(x) = \frac{H(x)}{\sum_i H(x_i)} \quad (\text{A.10})$$

where the values for x are shifted according to any offset between the two motif sequences. Values of $H(x)$ are buffered by a minimum value of 1e-45 to prevent extreme KL divergence values (i.e., 0 or infinity). The KL divergence V for two distributions $p_{w_1}(x)$ and $p_{w_2}(x)$ is calculated to be

$$V = \sum_x p_{w_1}(x) \cdot \log \left[\frac{p_{w_1}(x)}{p_{w_2}(x)} \right] \quad (\text{A.11})$$

The V -value threshold was set to 0.2 during our analyses; motif-pairs with similar sequences were clustered if their KL divergence fell below this threshold.

A.6 Consensus sequence determination and known *cis*-regulatory motif comparisons

Motif clusters are condensed into a single consensus sequence according to the criteria derived from [80] and [22]. Each aligned site is assigned a single residue consensus if it comprises 50% of the aligned 6mers and occurs at least twice as frequently as every other nucleotide. Double nucleotide degeneracy is applied to sites for which the two residues comprise 75% of the cases, with neither residue matching the criteria for a single site consensus. Sites not matching the criteria for either single or double nucleotide degeneracy are considered completely degenerate; triple degeneracy is not considered. During our analyses, comparisons to known regulatory elements in TRANSFAC v11.3 [80] were conducted using STAMP [77]; only binding motifs found in humans were considered.

Appendix B

MRF methodology

B.1 Statistical framework of the MRF model

The MRF model provides a measure of inter-motif distance preferences between two motifs. For any pair of motifs w and v , we define an MRF $f_{w|v}(x)$ to be the underlying frequency of w to occur exactly x bp from v . Thus we set:

$$f_{w|v}(x) = \Pr(U_{l_w}(x+i) = w | U_{l_v}(i) = v) \quad (\text{B.1})$$

We note that the position of v , given by i , defines the position $x = 0$. The function $f_{w|v}(x)$ is independent of i ; i.e., MRFs are defined as a conditional, rather than joint, probability.

Like MLFs, MRFs are modeled as the sum of a background function $C(x)$ and a signal function $H(x)$, as in Eq A.2. However, both the form of both $C(x)$ and $H(x)$ are extended from that of an MLF. The background function $C(x)$ of an MRF still represents the background frequency of motif occurrence for w , although in this case this background frequency is conditioned upon the distance from a known occurrence of v (see below). The form of the signal function $H(x)$ is also extended from that of an MLF. Specifically, the signal function of an MRF is modeled in order

to incorporate multiple peaks into the model, and is therefore modeled using a linear combination of unnormalized Gaussian terms:

$$H(x) = \sum_{j=1}^M a_j \cdot \exp \left[-\frac{(x - \mu_j)^2}{2\sigma_j^2} \right] \quad (\text{B.2})$$

where M represents the number of Gaussian terms (overrepresentation peaks). The number of overrepresentation peaks M is set to zero for motif-pairs not exhibiting spatial preferences, while motif-pairs exhibiting spatial preferences are modeled using one or more Gaussian terms ($M > 0$). For pairs of motifs exhibiting spatial preferences, the value for M reflects the number of inter-motif distances at which the pair of motifs tend to co-occur preferentially.

B.2 Background functions

The background function $C_{w|v}(x)$ of an MRF is estimated using a ‘prototype background function’ $c_{w|v}(x)$ in a similar as an MLF. In this case, however, $c_{w|v}(x)$ represents the expected probability for motif w to occur x bp away from motif v . This expected frequency is estimated according to the background functions of each individual motif, i.e., the MLF background functions of w and v : $C_w(x)$ and $C_v(x)$. Estimating the prototype background function first involves obtaining these MLF background functions as described in the previous chapter. We then assume the two motifs will occur randomly with respect to each other, and estimate $c_{w|v}(x)$ using the conditional probability provided in Eq B.1:

$$c_{w|v}(x) = \frac{\Pr(U_{l_w}(x+i) = w, U_{l_v}(i) = v)}{\Pr(U_{l_v}(i) = v)} = \frac{\int_i C_w(x+i)C_v(i)di}{\int_i C_v(i)di} \quad (\text{B.3})$$

The background function $C_{w|v}(x)$ is ultimately derived similarly to that of an MLF. Again, we set $C(x)$ to be $C_{w|v}(x) = b + d \cdot c_{w|v}(x)$, where b and d are free parameters in

the model. This allows the background frequency of occurrence to shift and stretch vertically with respect to its expected value $c_{w|v}(x)$.

B.3 MRF parameter estimation

To estimate the values of the parameters of a given an MRF $f_{w|v}(x)$, we maximize the log-likelihood of the data D given the model f . This is done in a similar manner as with an MLF, where the data D is set to be $D = \{\langle x_1, z_1 \rangle, \dots, \langle x_n, z_n \rangle\}$. The log-likelihood for a single data point is given by the binomial distribution:

$$L(\langle x_i, z_i \rangle; \theta_f) \equiv z_i \cdot \log[f(x)] + (N_i - z_i) \cdot \log[1 - f(x)] \quad (\text{B.4})$$

This form is similar to that used for MLF estimation, although for MRFs, the value N_i represents the maximum possible value for the number of motif occurrence z_i at the given position (x_i). This value differs according to position, as the motif w cannot co-occur with motif v when the latter is located on the edge of the promoter sequence window (in our application of the MRF model, the promoter sequence window comprises 500 bp prior to and 100 bp after the TSS). Thus, the value for N_i decreases as the absolute value of x_i increases. The total log-likelihood of the data given f is then given by the sum of the log-likelihoods across all positions. In our application, we choose the data points within $|x| < 150$; additionally, we buffer the values of x to be greater than 20 bp.

B.4 MRF model selection

The number of Gaussian terms M within the signal function $H(x)$ within an MRF (Eq B.2) are determined in an incremental fashion. We begin by comparing the model where $M = 0$ to the model where $M = 1$. In each case, we determine the MRF parameter values by maximizing the log-likelihood of the data given each of these two models. We then conduct a likelihood ratio test (F-test) in an identical fashion

as with the MLF model. Significant p -values then suggest that the two motifs share spatial preferences in relation to each other. In cases for which this first comparison produces a significant p -value, we then test for the existence of a second significant overrepresentation peak, comparing the model where $M = 1$ to that in which $M = 2$. Model comparisons are continued, adding statistically significant overrepresentation peaks (i.e., incrementing the value of M) until the F-test produces a non-significant p -value. The final MRF is taken to be the last function that produces a significant p -value.

Appendix C

Derivation of the intergenic background model Z-score function

The expected and variance terms of $T_{ij} - T_{ji}$ in Eq 5.1 are estimated according to a background frequency (p_{ij}) of co-occurrence. Consistent with our null hypothesis that nucleotide preferences are identical in the two species and that the background frequencies of co-occurrence can be estimated using the occurrences within the intergenic sequences, we estimate p_{ij} to be the average of the proportions b_{ij}/N_b and b_{ji}/N_b . So, we have that $\hat{p}_{ij} = (b_{ij} + b_{ji})/(2N_b)$. Our null hypothesis gives $\mathbb{E}[T_{ij}] = \mathbb{E}[T_{ji}] = N_t p_{ij}$. Because $\mathbb{E}[T_{ij}] = \mathbb{E}[T_{ji}]$, the term $\mathbb{E}[T_{ij} - T_{ji}]$ is zero and vanishes from the numerator.

The variance term $Var [T_{ij} - T_{ji}]$ is derived in the following way. First, the total variance $Var [T_{ij} - T_{ji}]$ is given by

$$Var [T_{ij} - T_{ji}] = Var [T_{ij}] + Var [T_{ji}] + 2Cov [T_{ij}, -T_{ji}] \quad (\text{C.1})$$

We see that the variance of the sum, rather than the difference, of T_{ij} and T_{ji} (i.e., $Var [T_{ij} + T_{ji}]$) can be written in two different ways:

$$Var [T_{ij} + T_{ji}] = Var [T_{ij}] + Var [T_{ji}] + 2Cov [T_{ij}, T_{ji}] \quad (\text{C.2})$$

and

$$Var [T_{ij} + T_{ji}] = N_t(2p_{ij})(1 - 2p_{ij}) \quad (C.3)$$

Eq C.3 can therefore be written as $Var [T_{ij} + T_{ji}] = N_t(2p_{ij} - 4p_{ij}^2)$, while Eq C.2 is $Var [T_{ij} + T_{ji}] = 2N_tp_{ij}(1 - p_{ij}) + 2Cov [T_{ij}, T_{ji}]$. Considering both equations and solving for $2Cov [T_{ij}, T_{ji}]$, we have

$$2Cov [T_{ij}, T_{ji}] = -2N_tp_{ij}^2 \quad (C.4)$$

But we note that $2Cov [T_{ij}, -T_{ji}] = -2Cov [T_{ij}, T_{ji}]$, and thus Eq C.1 becomes

$$Var [T_{ij} - T_{ji}] = 2N_tp_{ij}(1 - p_{ij}) + 2N_tp_{ij}^2 \quad (C.5)$$

Substituting the estimated value of p_{ij} to be the observed value $\hat{p}_{ij} = (b_{ij} + b_{ji})/(2N_b)$, our null hypothesis estimates $Var [T_{ij} - T_{ji}]$ to be $2N_t\hat{p}_{ij} = (N_t/N_b)(b_{ij} + b_{ji})$, as shown in Equation 5.2.

Bibliography

- [1] A. Abzhanov, M. Protas, B. R. Grant, P. R. Grant, and C. J. Tabin. Bmp4 and morphological variation of beaks in darwin's finches. *Science*, 305:1462–1465, 2004.
- [2] S. Aerts, P. Van Loo, G. Thijs, Y. Moreau, and B. De Moor. Computational detection of *cis*-regulatory modules. *Bioinformatics*, 19:ii5–ii14, 2003.
- [3] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51–80, 1995.
- [4] P. A. Beachy, J. Varkey, K. E. Young, D. P. von Kessler, B. I. Sun, and S. C. Ekker. Cooperative binding of an *Ultrabithorax* homeodomain protein to nearby and distant DNA sites. *Mol. Cell Biol.*, 13(11):6941–6956, 1993.
- [5] M. A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117:185–198, 2004.
- [6] R. Benfante, R. A. Antonini, M. Vaccari, A. Flora, F. Chen, F. Clementi, and D. Fornasari. The expression of the human neuronal alpha3 Na⁺, K⁺-ATPase subunit gene is regulated by the activity of the Sp1 and NF-Y transcription factors. *Biochem. J.*, 386:63–72, 2005.
- [7] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res.*, 38:D46–D51, 2010.
- [8] J. Berg, S. William, and M. Lassig. Adaptive evolution of transcription factor binding sites. *BMC Evol. Biol.*, 4:42, 2004.
- [9] M. F. Berger, G. Badis, A. R. Gehrke, S. Talukder, A. A. Philippakis, L. Pena-Castillo, and *et al.* Variation in homeodomain dna binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133:1266–1276, 2008.

- [10] M. D. Biggin and W. McGinnis. Regulation of segmentation and segmental identity by *Drosophila* homeoproteins: the role of DNA binding in functional activity and specificity. *Development*, 124:4425–4433, 1997.
- [11] A. P. Bird. CpG-rich islands and the function of dna methylation. *Nature*, 321:209–213, 1986.
- [12] A. R. Borneman, T. A. Gianoulis, Z. D. Zhang, H. Yu, J. Rozowsky, M. R. Seringhaus, L. Y. Wang, M. Gerstein, and M. Snyder. Divergence of transcription factor binding sites across related yeast species. *Science*, 317:815–819, 2007.
- [13] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, 2008.
- [14] C. G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, 19(92):577–593, 1965.
- [15] P. Bucher. Weight matrix descriptions of four eukaryotic rna polymerase ii promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, 212:563–578, 1990.
- [16] M. L. Bulyk, A. M. McGuire, N. Masuda, and G. M. Church. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. *Genome Res.*, 14:201–208, 2004.
- [17] H. J. Bussemaker, H. Li, and E. D. Siggia. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *PNAS*, 97(18):10096–10100, 2000.
- [18] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nat. Genet.*, 27:167–171, 2001.
- [19] J. Carcamo, L. Buckbinder, and D. Reinberg. The initiator directs the assembly of a transcription factor iid-dependent transcription complex. *Proc. Natl. Acad. Sci.*, 88:8052–8056, 1991.
- [20] S. B. Carroll. Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, 101:577–580, 2000.
- [21] S. B. Carroll. Evo-devo and expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134:25–36, 2008.

- [22] D. R. Cavener. Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.*, 15(4):1353–1361, 1987.
- [23] J. Chai and A. S. Tarnawski. Serum response factor: discovery, biochemistry, biological roles and implications for tissue injury healing. *J. Physiol. Pharmacol.*, 53(2):147–157, 2002.
- [24] P. C. Champ, S. Maurice, J. M. Vargason, T. Camp, and P. S. Ho. Distributions of z-dna and nuclear factor i in human chromosome 22: a model for coupled transcriptional regulation. *Nucleic Acids Res.*, 32:6501–6510, 2004.
- [25] E. T. Chan, G. T. Quon, G. Chua, T. Babak, M. Trochesset, R. A. Zirngibl, and *et al.* Conservation of core gene expression in vertebrate tissues. *J. Biol.*, 8:33, 2009.
- [26] A. V. Chernatyskaya, L. Deleeuw, J. O. Trent, T. Brown, and A. N. Lane. Structural analysis of the dna target site and its interaction with mbp1. *Org. Biomol. Chem.*, 7:4981–4991, 2009.
- [27] P. Clifton, P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, 301:71–76, 2003.
- [28] D. M. Crothers, T. E. Haran, and J. G. Nadeau. Intrinsically bent DNA. *J. Biol. Chem.*, 265(13):7093–7096, 1990.
- [29] A. C. Davison. *Statistical Models*. Cambridge University Press, New York, 2003.
- [30] E. T. Dermitzakis and A. G. Clark. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, 19(7):1114–1121, 2002.
- [31] S. W. Doniger and J. C. Fay. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol.*, 3:e99, 2007.
- [32] T. A. Down, C. M. Bergman, J. Su, and T. J. P. Hubbard. Large-scale discovery of promoter motifs in *Drosophila melanogaster*. *PLoS Comput. Biol.*, 3(1):0095–0109, 2007.

- [33] R. Elkon, C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh. Genome-wide *in silico* identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.*, 13:773–780, 2003.
- [34] E. Emberly, N. Rajewsky, and E. D. Siggia. Conservation of regulatory elements between two species of drosophila. *BMC Bioinformatics*, 4:57, 2003.
- [35] I. Ezkurdia, L. Bartoli, P. Fariselli, R. Casadio, A. Valencia, and M. L. Tress. Progress and challenges in predicting protein-protein interaction sites. *Briefings in Bioinformatics*, 10(3):233–246, 2009.
- [36] G. Felsenfeld and M. Groudine. Controlling the double helix. *Nature*, 421:448–453, 2003.
- [37] S. Fisher, E. A. Grice, R. M. Vinton, S. L. Bessling, and A. S. McCallion. Conservation of ret regulatory function from human to zebrafish without sequence similarity. *Science*, 312:276–279, 2006.
- [38] P. C. FitzGerald, A. Shlykhtenko, A. A. Mir, and C. Vinson. Clustering of DNA sequences in human promoters. *Genome Res.*, 14:1562–1574, 2004.
- [39] P. C. FitzGerald, D. Sturgill, A. Shyakhtenko, B. Oliver, and C. Vinson. Comparative genomics of *Drosophila* and human core promoters. *Genome Biol.*, 7(7):R53.1–R53.22, 2006.
- [40] C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen. Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, 299:283–293, 2000.
- [41] R. Groisman, H. Masutani, M. P. Leibovitch, P. Robin, I. Soudant, D. Trouche, and A. Harel-Bellan. Physical interaction between the mitogen-responsive serum response factor and myogenic basic-helix-loop-helix proteins. *J. Biol. Chem.*, 271(9):5258–5264, 1996.
- [42] M. Hanlon, T. W. Sturgill, and L. Sealy. Erk2- and p90-rsk2-dependent pathways regulate the ccaat/enhancer-binding protein-beta interaction with serum response factor. *J. Mol. Chem.*, 276(42):38449–38456, 2001.
- [43] S. Hannenhalli and S. Levy. Predicting transcription factor synergism. *Nucleic Acids Res.*, 30(19):4278–4284, 2002.

- [44] E. E. Hare, B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen. Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation. *PLoS Genet.*, 4:e1000106, 2008.
- [45] R. E. Harrington. Dna curving and bending in protein-dna recognition. *Mol. Microbiol.*, 6:2549–2555, 1992.
- [46] R. E. Harrington. Dna curving and bending in protein-dna recognition. *Mol. Microbiol.*, 6:2549–2555, 1992.
- [47] S. B. Hedges, J. Dudley, and S. Kumar. Timetree: A public knowledge-base of divergences times among organisms. *Bioinformatics*, 22:2971–2972, 2006.
- [48] J. Hizver, H. Rozenberg, F. Forlow, D. Rabinovich, and Z. Shakked. DNA bending by an adenine-thymine tract and its role in gene regulation. *PNAS*, 98(15):8490–8495, 2001.
- [49] H. E. Hoekstra and J. A. Coyne. The locus of evolution: evo devo and the genetics of adaptation. *Evolution*, 61:995–1016, 2007.
- [50] K. Huang, J. M. Louis, L. Donaldson, F.-L. Lim, A. D. Sharrocks, and G. M. Clore. Solution structure of the mef2a-dna complex: structural basis for the modulation of dna binding and specificity by mads-box transcription factors. *EMBO J.*, 19(11):2615–2628, 2000.
- [51] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, 295:1205–1214, 2000.
- [52] S. Huntley, D. M. Baggott, A. T. Hamilton, M. Tran-Gyamfi, S. Yang, J. Kim, L. Gordon, E. Branscomb, and L. Stubbs. A comprehensive catalog of human krab-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res.*, 16:669–677, 2006.
- [53] I. Ioshikhes, E. N. Trifonov, and M. Q. Zhang. Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America*, 96:2891–2895, 1999.
- [54] S. Jones, P. van Heyningen, H. M. Berman, and J. M. Thornton. Protein-dna interactions: a structural analysis. *J. Mol. Biol.*, 287:877–896, 1999.

- [55] T. Juven-Gershon, J.-Y. Hsu, and J. T. Kadonaga. Perspectives on the rna polymerase ii core promoter. *Biochem. Soc. Trans.*, 34:1047–1050, 2006.
- [56] T. Juven-Gershon, J.-Y. Hsu, J. W. M. Theisen, and J. T. Kadonaga. The rna polymerase ii core promoter—the gateway to transcription. *Curr. Opin. Cell Biol.*, 20(3):253–259, 2008.
- [57] S. Karlin, C. Burge, and A. M. Campbell. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.*, 20:1363–1370, 1992.
- [58] D. Karolich, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Hausler, and W. J. Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, 32:D493–D496, 2004.
- [59] S. Keles, M. J. van der Laan, and C. Vulpe. Regulatory motif finding by logic regression. *Bioinformatics*, 20(16):2799–2811, 2004.
- [60] M. Kellis, N. Paterson, B. Birren, B. Berger, and E. S. Lander. Methods in comparative genomics: Genome correspondence, gene identification, and regulatory motif discovery. *J. Comput. Biol.*, 11(2-3):319–355, 2004.
- [61] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423:241–249, 2003.
- [62] S. Kellis, M. van der Laan, and M. B. Eisen. Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18(9):1167–1175, 2002.
- [63] T. K. Kerppola. Transcriptional cooperativity: bending over backwards and doing the flip. *Structure*, 6:549–554, 1998.
- [64] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188:107–116, 1975.
- [65] S. Kullback and R. A. Leibler. On information theory and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.
- [66] D. S. Latchman. *Eukaryotic transcription factors*. Elsevier Academic Press, Boston, 4 edition, 2004.

- [67] C. E. Lawrence and A. A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *PROTEINS: Structure, Function, and Genetics*, 7:41–51, 1990.
- [68] S. Levy and S. Hannenhalli. Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome*, 13:510–514, 2002.
- [69] B. Lewin. *Genes VII*. Oxford University Press, Oxford, 2000.
- [70] L. Li, A. S. Cheng, V. X. Jin, H. H. Paik, M. Fan, X. Li, W. Zhang, J. Robarge, C. Balch, R. V. Davuluri, S. Kim, T. H. Huang, and K. P. Nephew. A mixture model-based discriminate analysis for identifying ordered transcription factor binding site pairs in gene promoters directly regulated by estrogen receptor-alpha. *Bioinformatics*, 22(18):2210–2216, 2006.
- [71] N. Li and M. Tompa. Analysis of computational approaches for motif discovery. *Algorithms for Molecular Biol.*, 1:doi:10.1186/1748–7188–1–8, 2006.
- [72] X. Y. Li, R. Bourgon, D. Nix, D. A. Pollard, V. N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C. L. Luengo Hendricks, H. C. Chu, N. Ogawa, W. Inwood, V. Sementchenko, A. Beaton, R. Weiszmann, S. E. Celniker, D. W. Knowles, T. Gingeras, T. P. Speed, M. B. Eisen, and M. D. Biggin. Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. *PLoS Biol.*, 6(2):e27, 2008.
- [73] C. Liberati, A. di Silvio, S. Ottolenghi, and R. Mantovani. NF-Y binding to twin CCAAT boxes: role of Q-rich domains and histone fold helices. *J. Mol. Biol.*, 285(4):1441–1455, 1999.
- [74] P. Lonai and A. Orr-Urteger. Homeogenes in mammalian development and the evolution of the cranium and central nervous system. *FASEB J.*, 4(5):1436–1443, 1990.
- [75] X.-T. Ma and H.-X. Tang. Predicting polymerase ii core promoters by cooperating transcription factor binding sites in eukaryotic genes. *Acta Biochimica et Biophysica Sinica*, 36:250–258, 2004.
- [76] D. R. Maglott, K. S. Katz, H. Sicotte, and K. D. Pruitt. Ncbi’s locuslink and refseq. *Nucleic Acids Res.*, 28(1):126–128, 2000.
- [77] S. Mahony and P. V. Benos. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, 35:W253–W258, 2007.

- [78] R. S. Mann and S. K. Chan. Extra specificity from extradenticle: the partnership between hox and pbx/exd homeodomain proteins. *Trends Genet.*, 12:258–262, 1996.
- [79] E. Martinez, C. M. Chiang, H. Ge, and R. G. Roeder. Tata-binding protein associated factor(s) in tfiid function through initiator to direct basal transcription from a tata-less class ii promoter. *EMBO*, 13:3115–3126, 1994.
- [80] V. Matys, E. Fricke, R. Geffers, E. Gobling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Micheal, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31(1):374–378, 2003.
- [81] M. Megraw, F. Pereira, S. T. Jensen, U. Ohler, and A. G. Hatzigeorgiou. A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res.*, 19:644–656, 2009.
- [82] J. M. Miano, X. Long, and K. Fujiwara. Serum response factor: master regulator of the actin cytoskeleton and contractile apparatus. *Am. J. Physiol. Cell Physiol.*, 292:C70–C81, 2007.
- [83] T. S. Mikkelsen, M. J. Wakefield, B. Aken, C. T. Amemiya, J. L. Chang, and *et al.* Genome of the marsupial monodelphis domestica reveals innovation in non-coding sequences. *Nature*, 447(7141):167–177, 2007.
- [84] C. T. Miller, S. Beleza, A. A. Pollen, D. Schluter, R. Kittles, M. D. Shriver, and D. M. Kingsley. Cis-regulatory changes in kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell*, pages 1179–1189, 2007.
- [85] W. Miller, K. Rosenbloom, R. C. Hardison, M. Hou, J. Taylor, B. Raney, and *et al.* 28-way vertebrate alignment and conservation track in the ucsc genome browser. *Genome Res.*, 17(12):1797–1808, 2007.
- [86] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.
- [87] L. H. Naylor and E. M. Clark. d(tg).d(ca) sequences upstream of the rat prolactin gene form z-dna and inhibit gene transcription. *Nucleic Acids Res.*, 18(6):1595–1601, 1990.

- [88] K. Noto and M. Craven. Learning probabilistic models of *cis*-regulatory modules that represent logical and spatial aspects. *Bioinformatics*, 23:e156–e162, 2006.
- [89] K. Nowick and L. Stubbs. Lineage-specific transcription factors and the evolution of gene regulatory networks. *Brief. Funct. Genomic Proteomic*, 1:65–78, 2010.
- [90] F. D. Nunes, F. C. de Almeida, R. Tucci, and S. C. de Sousa. Homeobox genes: a molecular link between development and cancer. *Pesqui Odontol Bras.*, 17(1):94–98, 2003.
- [91] D. T. Odom, R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, and *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genet.*, 39:730–732, 2007.
- [92] Y. Pan and R. Nussinov. p53-induced dna bending: the interplay between p53-dna and p53-p53 interactions. *J. Phys. Chem.*, 112:6716–6724, 2008.
- [93] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole. Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, 32:W199–W202, 2004.
- [94] A. G. Pederson, B. Pierre, Y. Chauvin, and S. Brunak. The biology of eukaryotic promoter prediction: a review. *Comput. Chem.*, 23(3-4):191–207, 1999.
- [95] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, 29:153–159, 2001.
- [96] R. Piskol and W. Stephan. Analyzing the evolution of rna secondary structures in vertebrate introns using kimura’s model of compensatory fitness interactions. *Mol. Biol. Evol.*, 25:2483–2492, 2008.
- [97] C. Plessy, T. Dickmeis, F. Chalmel, and U. Strahle. Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends Genet.*, 21:207–210, 2005.
- [98] J. Ponjavic, B. Lenhard, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and A. Sandelin. Transcriptional and structural impact of tata-initiation site spacing in mammalian promoters. *Genome Biol.*, 7(8):R78, 2006.
- [99] M. J. Potthoff and E. N. Olson. Mef2: a central regulator of diverse developmental programs. *Development*, 134:4131–4140, 2007.

- [100] P. Prabakaran, J. G. Siebers, S. Ahmad, and M. M. Gromiha. Classification of protein-dna complexes based on structural descriptors. *Structure*, 14:1355–1367, 2006.
- [101] K. D. Pruitt and D. R. Maglott. Refseq and locuslink: Ncbi gene-centered resources. *Nucleic Acids Res.*, 29(1):137–140, 2001.
- [102] K. D. Pruitt, T. Tatusova, and D. R. Maglott. Ncbi reference sequence (ref-seq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 33:D501–D504, 2005.
- [103] A. Rich, A. Nordheim, and A. H. Wang. The chemistry of biology of left-handed dna. *Annu. Rev. Biochem.*, 53:791–846, 1984.
- [104] R. G. Roeder. The role of general initiation factors in transcription by rna polymerase ii. *TIBS*, 21:327–335, 1996.
- [105] F. P. Roth, P. W. Hughes, J. D. Estep, and G. M Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, 26:939–945, 1998.
- [106] S. Rothenburg, F. Koch-Nolte, A. Rich, and F. Haag. A polymorphic dinucleotide repeat in the rat nucleolin gene forms z-dna and inhibits promoter activity. *PNAS*, 98(16):8985–8990, 2001.
- [107] A. Sandelin, P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki, and D. A. Hume. Mammalian rna polymerase ii core promoters: insights from genome-wide studies. *Nature*, 8:424–436, 2007.
- [108] L. K. Savinkova, M. P. Ponomarenko, P. M. Ponomarenko, I. A. Drachkova, M. V. Lysova, T. V. Arshinova, and N. A. Kolchanov. Tata box polymorphisms in human gene promoters and associated hereditary pathologies. *Biochemistry*, 74(2):117–129, 2009.
- [109] G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans. Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, 266:141–162, 1996.
- [110] E. Segal and R. Sharan. A discriminative model for identifying spatial *cis*-regulatory modules. *J. Comput. Biol.*, 12(6):822–834, 2005.
- [111] E. Segal and J. Widom. Poly(da:dt) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.*, 19:65–71, 2009.

- [112] K. Senger, G. W. Armstrong, W. J. Rowell, J. M. Kwan, M. Markstein, and M. Levine. Immunity regulatory dnases share common organizational features in *Drosophila*. *Mol. Cell*, 13:19–32, 2004.
- [113] S. Sinha and M. Tompa. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, 31(13):3586–3588, 2003.
- [114] S. Sinha, E. van Nimwegen, and E. D. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19:i292–i301, 2003.
- [115] S. T. Smale. Core promoters: active contributors to combinatorial gene regulation. *Genes and Development*, 15:2503–2508, 2001.
- [116] D. L. Stern and V. Orgogozo. Is genetic evolution predictable? *Science*, 323:746–751, 2009.
- [117] B. D. Stuhl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403:41–45, 2000.
- [118] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nat. Genet.*, 22:281–285, 1999.
- [119] K. Tharakaraman, O. Bodenreider, D. Landsman, J. L. Spouge, and L. Marino-Ramirez. The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. *Nucleic Acids Res.*, 36(8):2777–2786, 2008.
- [120] The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437:69–87, 2005.
- [121] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–1122, 2001.
- [122] I. Tirosh, N. Barkai, and K. J. Verstrepen. Promoter architecture and the evolvability of gene expression. *J. of Biol.*, 9:95, 2009.
- [123] A. Tomovic and E. J. Oakeley. Computational structural analysis: multiple proteins bound to dna. *PLoS ONE*, 3:e3243, 2008.

- [124] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, and *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23:137–144, 2005.
- [125] J. Tung, A. Primus, A. J. Bouley, T. F. Steverson, S. C. Alberts, and G. A. Wray. Evolution of a malaria resistance gene in wild primates. *Nature*, 460:388–391, 2009.
- [126] J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842, 1998.
- [127] J. van Helden, A. F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, 28(8):1808–1818, 2000.
- [128] S. Vardhanabhuti, J. Wang, and S. Hannenhalli. Position and distance specificity are important determinants of *cis*-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.*, 35(10):3203–3213, 2007.
- [129] J. Wang and S. Hannenhalli. A mammalian promoter model links *cis*-elements to genetic networks. *Biochem. Biophys. Res. Commun.*, 347:166–177, 2006.
- [130] Q.-F. Wang, S. Prabhakar, Q. Wang, A. M. Moses, S. Chanan, M. Brown, M. B. Eisen, J.-F. Cheng, E. M. Rubin, and D. Boffelli. Primate-specific evolution of an *ldlr* enhancer. *Genome Biol.*, 7:R68, 2006.
- [131] W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, 278:167–181, 1998.
- [132] R. H. Waterston, K. Landblad-Toh, E. Birney, J. Rogers, J. F. Abril, and *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.
- [133] Wellcome Trust Sanger Institute.
- [134] A. G. West and A. D. Sharrocks. Mads-box transcription factors adopt alternative mechanisms for bending dna. *J. Mol. Biol.*, 286:1311–1323, 1999.
- [135] R. J. White. *Gene transcription: mechanisms and control*. Blackwell Science, Malden, Mass, 2001.

- [136] E. Whitelaw. The role of dna-binding proteins in differentiation and transformation. *J. Cell. Sci.*, 94(2):169–173, 1989.
- [137] S. R. Willy, R. Kobayashi, and J. T. Kadonaga. A basal transcription factor that activates or represses transcription. *Science*, 2000:982–985, 2000.
- [138] D. S. Wilson and C. Desplan. Structural basis of hox specificity. *Nat. Struct. Biol.*, 6:297–300, 1999.
- [139] M. D. Wilson, N. L. Barbosa-Morais, D. Schmidt, C. M. Conboy, L. Vanes, V. J. Tybulwicz, and *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science*, 322:434–438, 2008.
- [140] M. D. Wilson and D. T. Odom. Evolution of transcriptional control in mammals. *Curr. Opin. Genet. Div.*, 19:579–585, 2009.
- [141] A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, and *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, 3:e7, 2005.
- [142] N. A. Woychik and M. Hampsey. The rna polymerase ii machinery: Structure illuminates function. *Cell*, 108:453–463, 2002.
- [143] G. A. Wray. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*, 20:1377–1419, 2007.
- [144] G. A. Wray, M. W. Hahn, A. Ehab, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, 20(9):1377–1419, 2003.
- [145] H. Xi, Y. Yu, Y. Fu, J. Foley, A. Halees, and Z. Weng. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res.*, 17(6):798–806, 2007.
- [146] X. Xie, J. Liu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammal species. *Nature*, 434:338–345, 2005.
- [147] K. Yamada, H. Osawa, and D. K. Granner. Identification of proteins that interact with nf-ya. *FEBS. Lett.*, 460(1):41–45, 1999.

- [148] C. Yang, E. Bolotin, T. Jiang, F. M. Sladek, and E. Martinez. Prevalence of the initiator over the tata box in human and yeast genes and identification of dna motifs enriched in tata-less core promoters. *Gene*, 7:R78, 2007.
- [149] K. D. Yokoyama, U. Ohler, and G. A. Wray. Measuring spatial preferences at fine-scale resolution identifies known and novel *cis*-regulatory element candidates and functional motif-pair relationships. *Nucleic Acids Res.*, 37:e92, 2009.
- [150] C. Zhou, F. Zhou, and Y. Xu. Comparative analyses of distributions and functions of z-dna in arabidopsis and rice. *Genomics*, 93:383–391, 2009.
- [151] Q. Zhou and W. H. Wong. Cismodule: *de novo* discovery of *cis*-regulatory modules by hierarchical mixture modeling. *PNAS*, 101(33):12114–12119, 2004.
- [152] C. Zhu, F. E. Johansen, and R. Prywes. Interaction of atf6 and serum response factor. *Mol. Cell. Biol.*, 17(9):4957–4966, 1997.

Biography

Ken Daigoro Yokoyama was born in Saint Louis, MO on May 23, 1982. He received a high-school diploma (Spring of 2000) in at Jamesville-Dewitt High School in New York state. He continued his study at University of California, Berkeley, obtaining a MA degree in Pure Mathematics in Spring of 2004. Subsequently, he worked for eight months as a research assistant at the National Institute of Genetics in Mishima, Japan. In 2005, he became a graduate student at Duke University in the Computational Biology and Bioinformatics program, pursuing a PhD under his advisor Gregory A. Wray.

Fellowships

James B. Duke Fellowship (Fall 2005-Spring 2009)

National Library of Medicine (NLM) Research Training Grant (Summer 2010–present)

Publications

Yokoyama KD, Thorne JL, Wray GA. (In review) Coordinated genome-wide modifications within proximal promoter *cis*-regulatory elements during vertebrate

evolution.

Yokoyama KD, Ohler U, Wray GA. (2009) Measuring spatial preferences at fine-scale resolution identifies known and novel *cis*-regulatory element candidates and functional motif-pair relationships. *Nucleic Acids Res.* 37(13): e92.

Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature Genet.* 39(9): 1140-1144.