Genomic Studies of Motif Enrichment and Conservation in the Regulation of Gene Expression in the Brain

by

David Alan Harmin

B.A., Music and Physics, Wesleyan University (1976)
M.S., Physics, University of Chicago (1978)
Ph.D., Physics, University of Chicago (1981)

Submitted to the Harvard-MIT Division of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

Master of Science in Biomedical Informatics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008

Signature of Author .....
Harvard-MIT Division of Health Sciences and Technology
May 9, 2008

Certified by .....
Isaac S. Kohane, M.D., Ph.D.
Lawrence J. Henderson Associate Professor of Pediatrics and Health Sciences and Technology, HMS, CHMC
Thesis Supervisor

Accepted by .....
Martha L. Gray, Ph.D.
Edwin Hood Taplin Professor of Medical and Electrical Engineering,
Director, Harvard-MIT Division of Health Sciences and Technology

Genomic Studies of Motif Enrichment and Conservation in the Regulation of Gene Expression in the Brain

by

David Alan Harmin

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on May 9, 2008, in partial fulfillment of the
requirements for the degree of
Master of Science in Biomedical Informatics

Abstract

Several bioinformatic tools will be brought to bear in this thesis to identify specific genomic loci that serve as regulatory gateways of gene expression in brain. These "motifs" are short nucleotide patterns that occur in promoters and 5′ or 3′ untranslated regions of genes. Occurrences of motifs that function in eukaryotic genomes as, e.g., transcription factor binding sites or targets of RNA interference are assumed to lie at the nexus of several trends. Instances that are indeed regulatory and not just bits of random sequence should show evidence of actual binding of factors that have a significant effect on expression levels. Such motif instances are also expected to be significantly enriched (or de-enriched), compared to background, in the genes regulated by their binding factors and in brain structures most closely associated with these genes' functions. Finally, truly regulatory motif instances are likely to be highly conserved in orthologous genes across multiple genomes; i.e., conservation can be taken as a proxy for function. My research exploits these ideas by exploring genome-wide properties of motifs associated with the transcription factor family MEF2, some of whose members are known to play a role in synapse development. Data from chromatin immunoprecipitation and tiling-microarray (ChIP-on-chip) experiments [1] have isolated peaks of specific binding by MEF2 in developing rat brains. Conservation and enrichment of these sites are analyzed here for their association with functionality and variability of motifs in genes that have been shown to fall under the control of MEF2 in excitatory neurons. The relationships between regulatory motif content, motif functionality, and expression of neuronal genes investigated in this work can help elucidate how programs of gene expression are controlled—and hence how they might go awry—in the brain.

Thesis Supervisor: Isaac S. Kohane, M.D., Ph.D.
Title: Lawrence J. Henderson Associate Professor of Pediatrics and Health Sciences and Technology, HMS, CHMC

Acknowledgments

Switching careers from theoretical physics to biomedical informatics and neuroscience has been no easy feat. I am deeply grateful to my colleagues, collaborators, mentors, friends, and family for their unconditional encouragement and assistance through these changes in my life. Their belief in my potential to make a contribution to such a different, fascinating, and by comparison practical field constitutes flattery of the highest order. I could not have negotiated so many professional and personal shifts and challenges without them. Prof. Michael Cavagnero at the University of Kentucky Department of Physics & Astronomy was a generous chairman indeed to support my leave of absence to dabble in biological benchwork in Boston. Dr. T-K Kim, Steve Flavell, Dr. Janine Zieg and other members of the Greenberg lab in the F.M. Kirby Neurobiology Center at Children's Hospital Boston and Harvard Medical School have been steady friends, patient teachers, and magicians deluxe, performing eye-opening experiments at the drop of a hat and producing magnificent data for bioinformatic processing. Dr. Jesse Gray, the best collaborator I've ever had, is a biological rennaisance man, an idea factory, sharp, voracious, willing to try anything—including the impossible, which he somehow manages to tame into existence. I am indebted to my mentor Zak Kohane, a rennaisance king, for welcoming me into the Children's Hospital Informatics Progam with an NLM fellowship and all the opportunities that CHIP has to offer for interactions with biomedical informaticians at the top of the game. His leadership and advice for the whole group and as my advisor have gone well beyond my expectations. HST has been a marvelous home for me at MIT, the best. My experience at MIT has been nothing short of wonderful; thanks to all the educators and researchers there who boosted me into bioinformatic orbit. The National Library of Medicine provided generous support through a Kirschstein training grant. Absolutely none of this would have been possible without the lasting friendship, loyalty, suggestions, insights, and scientific wizardry of Prof. Mike Greenberg, for whom a finite catalog of thanks will have to suffice. Matt, Cal, Craig—my glorious guys, best friends, the children I'd always wanted, thank you for all your affirmations (and for letting me graduate first!). And above all, my most heartfelt gratitude to my amazing wife Karen, World's Best Cook Quilter Dulcimer Player—I am awed by your generosity, spirit, devotion, and belief in me. It was so worth waiting thirty-one years. And Bob.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Biomedical science in the genomic era has the opportunity to exploit insights afforded by genome-wide studies of how human cells attain and maintain their proper function in response to internal and external cues. Cells respond to their environment via signaling networks whose components are ultimately encoded in an organism's genetic instructions. The transcription of genes and the regulation of mechanisms that differentially control gene transcription in different types of cells are therefore critical to the specificity of cellular activity. Since most human genes are expressed in the brain [2], neuroscience in particular stands to benefit from a genomic approach to understanding the regulation of genes involved in the development, maturation, and function of neurons.

Bioinformatic tools can be brought to bear to identify specific genomic loci that serve as either pre- or post-transcriptional regulatory gateways of gene expression in the brain. These "motifs" are noncoding nucleotide consensus patterns of length 6–18 nucleotides (nt) that occur in noncoding regions closely associated with genes—i.e., nearby promoter regions and 3′ untranslated regions (3′UTRs) of genes, or more distal enhancers. Note that vertebrate promoters are usually assumed to lie upstream of (5′ to) the transcription start site (TSS), though this has been called into question [3]. The sequences found in promoters function as binding sites for transcription factors (TFs), while those in 3′UTRs are potential targets of miRNAs for mediating RNA interference (RNAi). There may be many sites in the genome that merely conform to motif patterns but do not normally bind functioning factors. Identification of instances in the genome that are actual sites of factor binding and functional activity will thus point to loci of key interactions in the regulation of gene transcription and transcript processing.

The research presented in this thesis covers one of several projects that I have

worked—described below—on that investigate genome-wide properties of motif distributions and their role in gene regulation. some of this work employed a set of over 1500 previously discovered, conserved motifs [4]. These "Xie motifs" were discovered in 4-kb–wide 5′ promoter regions and in 3′UTRs of over 17,000 human genes. They represent sites that are conserved across a 4-species genomic alignment (human, mouse, rat, dog) significantly more often than in nonregulatory regions. This set of motifs provides a reliable catalog of sequence patterns that are distinguished by their unusually high rate of conservation. However, even if sequence conservation is taken as strong evidence of functionality, most occurrences of these motifs are in fact not well conserved. There still remains the important challenge of ascertaining with reasonable confidence those particular consensus sites that are truly functional in neurons. Note that the issue here is not so much motif discovery (for which tools such as MEME exist [5]) as consensus site validation.

Motifs having instances in the human genome that function as either TF binding sites or targets of RNAi are assumed to exhibit several characteristics, which can be assessed bioinformatically:

- Instances that are indeed regulatory and not just bits of random sequence should show evidence of actual binding, e.g., by a TF as revealed through chromatin immunoprecipitation and microarray (ChIP-on-chip) experiments, and significantly high (or low) expression levels in particular brain regions compared to background levels.

- Such motif instances are also expected to be significantly enriched, compared to background, in the genes regulated by their binding factors and in brain structures most closely associated with these genes' functions. (Alternatively, motifs may be de-enriched in genes that should not be regulated by certain binding factors.)

- Finally, truly regulatory motif instances are likely to be highly conserved in orthologous genes across multiple genomes, implying that both these genes and these regulatory sequence(s) are confer functionality that has been fixed by evolution—i.e., conservation is taken as a proxy for function.

My research has investigated and quantified how these features collectively contribute to motif instance functionality. What follows is a brief description of three projects I have worked on, the last of which, on the transcription factor family MEF2 (*myocyte enhancer factor 2*), comprises the content of this thesis.

Deeper conservation of the Xie motifs and functional instances in human. This work was done in collaboration with Dr. Jesse M. Gray of the Michael E. Greenberg lab at Harvard Medical School and Children's Hospital Boston.[1] We further analyzed the Xie set of conserved motifs [4], which had been originally extracted from a 4-way alignment of the genomes `hg17-mm5-rn3-canFam1`, with the aim of identifying as many functional instances of these patterns as possible. This set covers most genes of the human genome. We matched all motif occurrences in promoter windows (TSS $\pm$ 4kb) and in UTRs for this set against a more extensive 17-way alignment of the human genome to 16 other vertebrate genomes created by UCSC [6]. (A 28-way alignment has become available more recently [7, 8], but we have not used it to re-analyze the Xie motifs.) The use of more genomes (including chimp, cow, opossum, chicken, ...) dispersed in evolutionary distance (... all the way down to frog, fugu, and zebrafish) provided a more reliable estimate of significant conservation for such short ($\sim$ 6 nt) nucleotide sequences [9]. The aim was to enhance the confidence with which functional regulatory elements could be pinpointed by allowing useful estimates and optimization of (low) false discovery rates.

Our analysis of the Xie motifs based on this deeper conservation resource has produced lists of individual motif instances that are likely to be functional with false discovery rates (FDR) of less than 25% or better. At the top of our highest-confidence list (FDR $\leq$ 5%) for 3′UTR motif instances are the polyadenylation (polyA) signal AATAAA and the motif TGTANATA for binding of *pumilio homolog 2* (PUM2)—a gene which is in fact conserved all the way down to *Drosophila* [10]. The former may be regarded as a positive control. But the ubiquity of well conserved PUM2 sites may be a more substantial discovery, considering the role of PUM2 in translational control; see, e.g., Ref. [11].

Correlation of Xie motif enrichment with gene expression and brain structure in mouse. The expression levels for over 20,000 mouse genes, obtained from high-resolution *in situ* hybridization data from each of 209 structures of the mouse brain, were made available through a collaboration between Dr. Gray and myself with the Allen Brain Institute [2]. Several hundred of the most highly expressing genes are currently being analyzed for their degree of enrichment (or de-enrichment) of the Xie set of highly conserved motifs in 17 major mouse brain regions. The ABA also made available to us a set of high-resolution three-dimensional expression data on a

[1] In this thesis I will mention my collaborators explicitly where appropriate. I will distinguish my specific contributions to our work by using the pronoun "I" and reserve "we" for parts for which I cannot take most of the credit. I will use the passive voice for work done by other groups.

$67 \times 41 \times 58$ grid (spacing 25 μm). (This data set was a more complete version of data accessible online from the Allen Brain Atlas [12].) In order to organize and enable us to visualize these data, I developed an algorithm for identifying regions of continuously correlated voxels. Without reference to mouse brain morphology as input, this method was able to capture major brain structures having consistent expression patterns. An outstanding project is to associate these spatial expression clusters with patterns of motif enrichment.

For the whole gene set of hybridization data taken from coronal or sagittal sections, those motifs that most highly correlate (positively or negatively) with genes expression patterns specific to each of the 209 structures will be identified via alternative machine-learning methods. I am planning to begin by applying Random Forests to these data [13, 14]. In this technique, many regression trees are generated randomly, where the branching in each tree is determined by how well the concentrations of a handful of random motifs (features) distinguish sets of resampled expression levels between structures (by entropy maximization) and where out-of-bag (nonsampled genes) expression levels are used as test sets.

Use of conservation in tandem with ChIP-on-chip to probe gene regulation by MEF2 in the rat brain. This project focuses on the particular family of transcription factors MEF2, which are highly expressed in brain and under the regulation of several calcium signalling pathways [15]. MEF2A and MEF2D have been shown to play a significant role in the regulation of synapse number in the developing central nervous system (CNS) [16]. Data from ChIP-on-chip and tiling-microarray experiments carried out by members of the Greenberg lab have been used to identify peaks of specific binding by the transcription factor MEF2D in developing rat brains [1].

I have analyzed conservation of these sites, their enrichment in the Mef2 binding peaks, and their variability, along with alternatives to the usual Mef2 consensus motif, for their influence on motif binding and functionality in genes whose expression is sensitive to MEF2 levels in excitatory neurons. This work is the main subject of this thesis, described in Chapters **2** and **3**.

Chapter 2

Use of Conservation in Tandem with ChIP-on-chip to Probe Gene Regulation by MEF2 in the Rat Brain

This chapter is organized as follows. Chromatin immunoprecipitation and microarray experiments to assay MEF2 binding are described in Sec. **2.1**. In Sec. **2.2** I survey canonical and variant consensus sites of MEF2 binding for their enrichment compared to random background levels. I analyze conservation properties of the Mef2 binding peaks in Sec. **2.3** and focus on the MEF2 and control motif consensus sites and appropriate ways to visualize their conservation in Sec. **2.4**.

I present conclusions from this work in Chapter **3**. Questions raised here that merit further investigation, as well as possible extensions to the current work, are collected in Chapter **4**.

2.1 MEF2 Peaks from ChIP and Tiled-Array Data

The following chromatin immunoprecipitation ("ChIP") and microarray experiments ("chip") were performed by Steven Flavell and Dr. Tae-Kyung Kim in the laboratory of M. E. Greenberg and are described in Ref. [1].

To develop mRNA profiles reflecting the effect of neuronal activity on the targeting of genes by MEF2, ChIP was performed on rat hippocampal neurons that had been harvested at E18, cultured 7–8 days *in vitro*, stimulated by treatment with KCl to induce membrane depolarization, and observed (1) before (i.e., at zero hours), one hour after, and six hours after stimulation. The neurons were then exposed to a Mef2D-specific antibody [16], which was cross-linked to DNA via paraformaldehyde treatment, nuclei were isolated, the DNA was sonicated to fragments of mean length $\sim$ 500 bp, crosslinking was reversed, and the surviving DNA fragments were purified. These ChIP samples and appropriate negative controls were amplified by PCR and ligated to linkers in preparation for microarray hybridization. Antibody efficiency and primer sets used for amplification were validated using quantitative-PCR and Western-blot analyses. Though the antibodies used were specific to Mef2D (*vs.* Mef2A), in the following we will refer to this transcription factor as simply *MEF2*.

Custom Nimblegen rat-genome tiling arrays were used with probes that covered the 308 genes described above, plus 40kb of "padding" outside both 5′ and 3′ ends of each gene. There was one 50-bp probe per 100–125 bp in these gene regions. In order to identify genomic loci of true MEF2 binding to the DNA of these neurons, we developed an algorithm for filtering probe intensities that optimized the heuristic for what would be considered a "Mef2 peak." To be conservative and maximize signal-to-noise, we considered only a minimal number of consecutive probes (physically adjacent in the rat genome) whose log2 intensity values exceeded a given percentage cutoff. Optimal criteria were found to be at least four consecutive probes at a 99% cutoff level (i.e., the 1% highest intensities). Three examples of identified Mef2 peaks are shown in Fig. **2-1** as UCSC Genome Browser tracks [17] along with their conservation tracks (described below in Sec. **2.3**). A total of 241 Mef2 peaks were found in the tiled gene regions. Note that my analysis here are based on the annotations provided with the UCSC *rn4* reference assembly [18] (based on RGSC v3.4). As a result, I will include only 296 of the 308 tiled genes: 296 are found in *rn4*, while 12 that are annotated instead in the Celera assembly are excluded as they have no matching annotation in *rn4*.

The point of the ChIP experiment is to isolate DNA fragments that are bound by

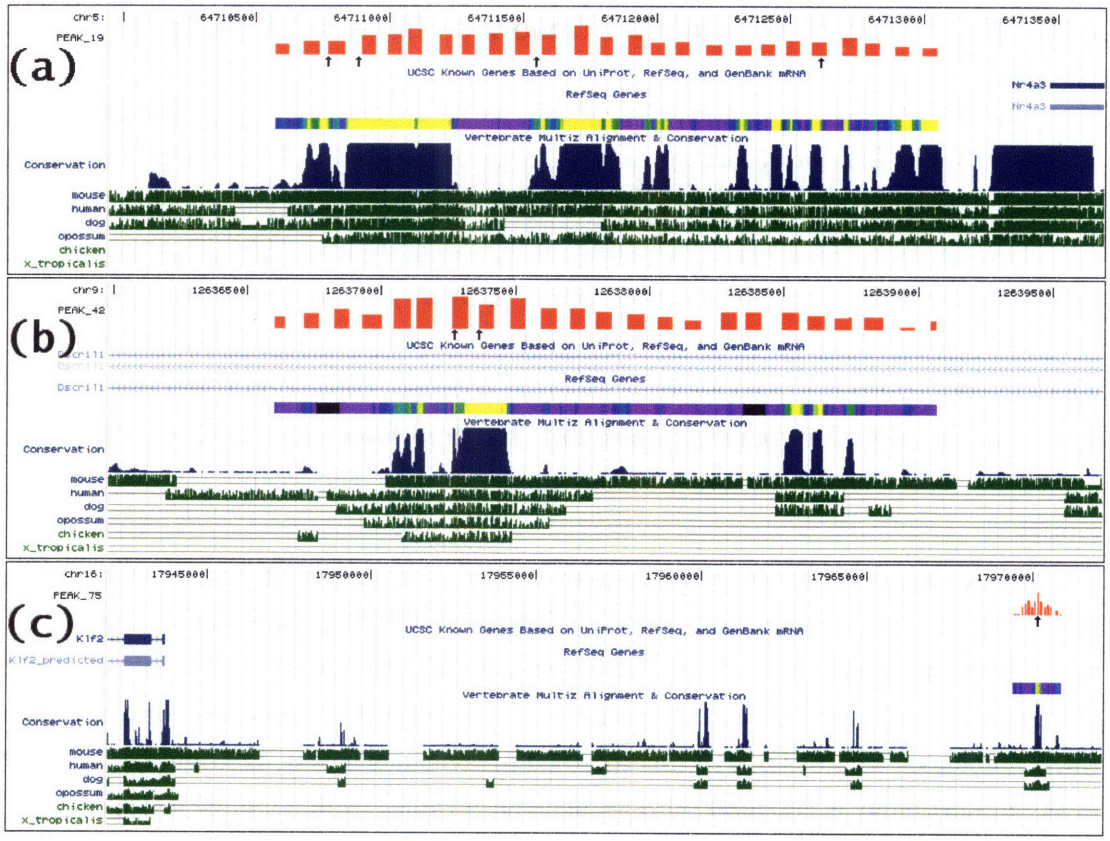

Figure 2-1. Three typical genomic sequences containing Mef2 peaks, delineated by consecutive microarray probes (red bars, heights proportional to normalized probe intensities); arrows, canonical MEF2 binding sites. **(a)** 3.7 kb on rat chr5 with a 2481-bp peak about 400 kb upstream of *Nr4a3*, an orphan nuclear receptor involved in regulating transcription; **(b)** 3.7 kb on rat chr9 with a 2470-bp peak within the last intron, 11 kb upstream of the last exon, of *Dscr1l1*, which binds to calcineurin and is involved with CNS development; **(c)** 30 kb on rat chr16 with a 1478-bp peak 26 kb upstream of *Klf2*, a transcription factor that regulates important aspects of vascular function. Conservation track values for peaks also shown color coded (high/*yellow* to low/*purple*) as in Sec. **2.3**; see Figs. **2-5**–**2-8**.

MEF2 in activity-regulated hippocampal neurons. Hence the 241 Mef2 peaks that were identified among the set of 296 rat genes can be expected with high probability to include functional binding sites of MEF2. We further expect such functional loci to be significantly conserved compared to the genome at large and even the tiled gene regions. Association of the transcription factor MEF2 with these loci may, however, be due to indirect binding of MEF2 to other components of a DNA-binding complex that controls transcription of the target genes. In fact, only 151 instances of the canonical MEF2 binding site are found in these peaks. It is therefore of interest to characterize any associations of these de facto Mef2 peaks with predicted MEF2 binding consensus sequences, possible variants of consensus sequences, and regions of high conservation.

2.2 Enrichment of MEF2 Consensus Sites and their Variants

One way to represent a consensus sequence is as a position-weight matrix, which assigns a probability of observing each base (A or C or G or T) at each position in the sequence. Here I will use the more compact standard notation, which represents each position by a single letter indicating either 1 definite base ({ACGT}), unresolved 2-base ({SWRYMK}) or 3-base ambiguities ({BDHV}), or no specificity ({N}). One can also represent per-position base frequencies visually as motif "logos" Ref. [19], which nicely capture a pattern's information content. Although some information about single nucleotide frequencies are lost, the notational convenience is preferable here so for simplicity I will stick to the ambiguated-letters scheme for representing motifs.

The canonical MEF2 consensus sequence has been inferred from SELEX-type experiments [20, 21]. The consensus for MEF2 is YTAWWWWTAR, which I will refer to as the motif *Mef2*. It is palindromic and AT-rich (W={AT}), with only the first and last bases allowing C or G. Note for future reference that the CG content of the rat genome as a whole is `0.418`, for the tiled array genes it is `0.428`, and in the Mef2 peaks in particular it is `0.489`. (I obtained these values from a simple census of base pairs in the relevant regions of the *rn4* reference assembly.)

The *Mef2* motif can be matched by $1^4 \times 2^6 = 64$ specific sequences of the 4 bases {ACGT}. An infinite random sequence of bases with 50% CG content would therefore be expected to have an occurrence of the motif with frequency $64/4^{10} = 2^{-14}$, or at 1 in every $2^{14} = 16,384$ bases on the average. Note that if YTAWWWWTAR were not palindromic, that this frequency would be approximately doubled (to 1 in 8,192) because each specific sequence could be found on either DNA strand. (See Appendix **A** for details.) More importantly, deviations from 50% CG can significantly alter the expected background rate of motif matches. For example, if the *Mef2* motif occurred randomly in the rat genome and if the genome had no structure, it would be found once every `4,862` bases on average. The details of this kind of accounting are spelled out in greater detail in Appendix **A.2**.

2.2.1 MEF2 Consensus Sites

The frequency with which 10-bp windows of the rat genome, or subsets of the rat genome, actually match the *Mef2* consensus can be inferred from a simple search of the published reference genome. I downloaded the complete *rn4* assembly of the rat genome from the UCSC Web site [22]. This included over 258Mb of sequence on

20 autosomes and 1 sex chromosome. To mine this data, I wrote a set of Perl programs that comprise a pipeline for finding motif instances (loci and sequence) in both genomic DNA and DNA subsequences, and for extracting sequence and conservation data from given loci; see Appendix **B** for details. Note that some of the enrichment and conservation characteristics described here were obtained by sampling only rat chromosome 2 ("chr2"); checks on whether it is indeed representative of the whole genome are mentioned where appropriate.

Based on the CG fraction of 0.418 for the rat genome as a whole, the expected background rate of 1 instances of YTAWWWWTAR per 4,862 bases equals 53,109 ± 230, where the "error" here quotes 1 standard deviation assuming Poisson statistics (see Appendix **A.1** and **A.2**). But the number of *Mef2* sites actually I found in the whole genome equals 68,779, a 1.3-fold enrichment. The collective region of all the tiled genes comprises 31.3Mb of sequence—about 12.1% of the rat genome—and is only slightly less AT-rich. Here the random background rate for this motif is 1 in 5,585, implying 5,606 ± 75 expected instances. The number of discovered *Mef2* instances is 7,535, again about a 1.3-fold enrichment. So this motif crops up fairly frequently in rat, but about as often within genes as without.

The situation in the *Mef2* peaks identified by the ChIP-chip experiments is dramatically different. Totalling just under 200kb, this region is relatively CG-rich ($p_{CG} = 0.489$) and the random background rate of YTAWWWWTAR occurrences drop to 1 in 13,766. The expected number of *Mef2* instances in the peaks is then only 14 ± 4, which seems low for sequences that were identified as bound by MEF2. *In fact, I found **151** instances of the canonical Mef2 motif in the Mef2 peaks' genomic subsequences.* This represents a *10.4-fold enrichment* over background in the peaks, or about *36 standard deviations* above what random sequence would deliver. I'll demonstrate in Sec. **2.4** another way in which this represents an unusually high level of significance. The 151 *Mef2* instances are located in 128 (53%) of the 241 Mef2 peaks. In total, 113 peaks had no consensus sites, while 110 had 1 sites, 14 had 2, 3 had 3, and 1 peak had 4 sites. The latter, shown in Fig. **2-1(a)**, lies just upstream of the transcription factor *Nr4a3* (nuclear receptor subfamily 4, group A, member 3).

There are two points of particular biological relevance here. The fact that barely half of the Mef2 peaks contained canonical MEF2 binding sites is indication that an appreciable amount of the interaction of MEF2 with chromatin takes place through *indirect binding*, as noted above. That is, MEF2 may frequently interact with DNA an as yet unspecified protein complex. In addition, it turns out that 64% of the binding sites identified through the ChIP-chip experiments [1] were 2.5kb or farther

from any gene's transcription start sites. Combined with the fact that the non-control probed genes were already known to be targets of MEF2 regulation, this suggests that both proximal and distal binding of MEF2 are important for its regulation of its target genes.

2.2.2 MEF2 Consensus Variants

Binding site motifs for transcription factors can also be inferred directly from genomic sequence. The assumption is that any sequence that has already been associated with a factor's binding, such as the Mef2 peaks here, should be enriched compared to an 8^{th}-order background of random sequence in whatever particular motifs serve as the true binding site. A popular program for discovering short, enriched motifs in DNA sequence is MEME [23, 5] (*M*ultiple *E*xpectation-Maximization for *M*otif *E*licitation). After supplying this tool with background information appropriate to the peaks' base composition, we elicited 10bp or 11bp motifs according to a few different criteria. In every case the Mef2 peak sequences were prefiltered to avoid regions of especially low evolutionary conservation by considering only those bases with phastCons scores $\geq$ 0.05 (see the Sec. **2.3** on conservation scores). Further distinctions were made based on the presence or absence of *Mef2* sites within the 1%-cutoff Mef2 peaks, yielding these four MEF2 consensus variants (*Mef2* itself is "number 0." on this list):

1. First, MEME was given only Mef2 peaks that contained at least one occurrence of the canonical binding site YTAWWWWTAR. The most significant motif returned by MEME was HWAWAWWWAR, which I will refer to as *Mef2-var1*. This pattern is relatively permissive, in the sense that it would be matched by random sequence at 1 in 2,376 sites (*vs.* 1 in 13,766 expected for *Mef2*). Strictly speaking, MEME didn't return the *Mef2* motif *per se*. But noting that $\mathsf{H} = \{\mathsf{A}, \mathsf{C}, \mathsf{T}\} \supset \mathsf{Y}$, *Mef2-var1* is essentially just a more lenient version of *Mef2*.

2–3. Second, MEME was given only Mef2 peaks that did *not* contain any occurrences of the canonical binding site YTAWWWWTAR. For these sequences MEME returned two motifs: KBYTDTTTWDD, called *Mef2-var2* here, and DRTWWTTWTAR, called *Mef2-var3*. These are, respectively, as permissive as *Mef2-var1* (1 in 2,861 for *Mef2-var2*) and less so than *Mef2* itself (1 in 18,221 for *Mef2-var3*).

4. Finally, MEME was given all Mef2 peaks, irrespective of whether any instances

Instances in rn4 Chromosome 2 (258,207,540 bp; $p_{CG} = 0.418$)							
***Mef2*-like Motifs**		**Expected Instances**			**Found Instances**		
Name	***Consensus***	***1 in (bp)***	#	$\pm\sigma$	#	z	×
Mef2	YTAWWWWTAR	4,861.8	53,109	230	67,878	64.1	1.3
Mef2-var1	HWAWAWWWAR	799.9	322,800	568	479,635	276.0	1.5
Mef2-var2	KBYTDTTTWDD	1,365.7	189,066	435	376,746	431.6	2.0
Mef2-var3	DRTWWTTWTAR	6,146.4	42,010	205	62,466	99.8	1.5
Mef2-HiSco	BTWTWTHWDDH	370.8	696,353	834	882,992	223.7	1.3
Mef2-Perm	WYWAATRWTW	2,430.9	106,219	326	127,800	66.2	1.2
Mef2-Rdm1	AKCTWWAGMT	37,700.7	6,849	83	6,471	-4.6	0.9
CREB	TGACGTMD	6,087.7	42,415	206	6,429	-174.7	0.2
Instances in Tiled-Array Genes (31,309,615 bp; $p_{CG} = 0.428$)							
***Mef2*-like Motifs**		**Expected Instances**			**Found Instances**		
Name	***Consensus***	***1 in (bp)***	#	$\pm\sigma$	#	z	×
Mef2	YTAWWWWTAR	5,584.9	5,606	75	7,443	24.5	1.3
Mef2-var1	HWAWAWWWAR	925.0	33,848	184	55,545	117.9	1.6
Mef2-var2	KBYTDTTTWDD	1,507.3	20,772	144	47,171	183.2	2.3
Mef2-var3	DRTWWTTWTAR	7,105.4	4,406	66	6,903	37.6	1.6
Mef2-HiSco	BTWTWTHWDDH	419.1	74,707	273	98,620	87.5	1.3
Mef2-Perm	WYWAATRWTW	2,792.4	11,212	106	13,157	18.4	1.2
Mef2-Rdm1	AKCTWWAGMT	39,900.3	785	28	656	-4.6	0.8
CREB	TGACGTMD	6,017.9	5,203	72	1,140	-56.3	0.2
Instances in MEF2 Peaks (199,498 bp; $p_{CG} = 0.489$)							
***Mef2*-like Motifs**		**Expected Instances**			**Found Instances**		
Name	***Consensus***	***1 in (bp)***	#	$\pm\sigma$	#	z	×
Mef2	YTAWWWWTAR	13,766.2	14	4	151	35.9	10.4
Mef2-var1	HWAWAWWWAR	2,375.9	84	9	341	28.1	4.1
Mef2-var2	KBYTDTTTWDD	2,860.7	70	8	287	26.0	4.1
Mef2-var3	DRTWWTTWTAR	18,221.3	11	3	53	12.7	4.8
Mef2-HiSco	BTWTWTHWDDH	926.3	215	15	549	22.7	2.5
Mef2-Perm	WYWAATRWTW	6,883.1	29	5	51	4.1	1.8
Mef2-Rdm1	AKCTWWAGMT	60,130.8	3	2	5	0.9	~ 1
CREB	TGACGTMD	5,918.9	34	6	24	-1.7	0.7

Table 2.1. Expected and Found Instances of *Mef2*-like Motifs in 3 rat genomic regions (with total length in bp and CG fraction p_{CG}). Expected: 1 in (bp), inverse average frequency of each motif for random sequence; $\# \pm \sigma$, expected number $\pm$ 1 standard deviation. Found: #, number of nonoverlapping instances in rn4 genome; z-score of found instances based on expected number; ×, fold enrichment over expected number.

of YTAWWWWTAR were present. The top returned motif, BTWTWTHWDDH, was the most general sequence returned by MEME. It is particularly permissive, occurring from 13 to 15 times more frequently than *Mef2* itself. It's first 10 of 11 bases include most sequences matched by *Mef2*, plus many more ($1^3 \times 2^5 \times 3^5 = 7{,}712$ possible matches in total, as opposed to *Mef2*'s 64). Its generality is of course offset by its nonspecificity.

Results for expected and found instances of YTAWWWWTAR and the above variants are collected in Table **2.1** for the three genomic regions of interest: (1) genome-wide, with the entire rn4 chromosome 2 used as a sample; (2) the 296 genes tiled on the array (including 10kb padding on both ends, see above); and (3) the Mef2 peaks. The expected number of null-hypothesis instances are simply based on sampling random sequences of the given total length with the indicated CG fraction. I obtained the found instances by counting nonoverlapping matches to the motif patterns in the relevant subsets of the rat reference genome. (Allowing matches to overlap approximately doubled the number of counted instances for the especially promiscuous motifs *Mef2-HiSco*, *Mef2-var1*, and *Mef2-var2*, while the other motif counts rose by only a few percent at most. These "redundant" occurrences reflected mostly AT-rich repetitions.) Poissonian standard deviations and z-scores follow the rationale explained in Appendix **A.2**. (The `perl` scripts I wrote are described in Appendix **B**.)

Statistics for 3 additional motifs employed as controls are shown in Table **2.1**:

5. A single random permutation WYWAATRWTW of the cannonical *Mef2* motif YTAWWWWTAR, which I call *Mef2-Perm*, is included to query whether just the particular bases and ambiguities of *Mef2* (2 A's, 1 R, 2 T's, 4 W's, 1 Y), rather than their particular order, accounts for the great enrichment of target sequences of MEF2. The 1.3-fold enrichment of *Mef2-Perm* is seen to be comparable to *Mef2* genome-wide and among the tiled genes. In the peaks, however, though the permuted motif is enhanced further to about 2× background, it is nowhere near the ≈ 10-fold enrichment of *Mef2*. Random permutations of the canonical motif are analyzed further in Sec. **2.4.3**.

6. Along the same lines, I created a "random" motif AKCTWWAGMT, called *Mef2-Rdm1*, with approximately the same AT content as YTAWWWWTAR as well as the same information content in its 10-bp pattern—i.e., 4 single-base positions plus 6 double-base positions ($4 \cdot 2 + 6 \cdot 1 = 14$ bits of information to fully specify the pattern)—but with the bases ambiguated in a scrambled order.

This should control for the *Mef2* motif's affinity for randomly matching 64 out of every 4^{10} 10-bp pattern with the same average AT content but without regard to two-base ambiguities and their positions in the pattern. *Mef2-Rdm1* is in fact slighly de-enriched compared to the expected background but not much different from the random rate.

7. The last motif, TGACGTMD, represents target sequences of the transcription factor *CREB*, which is also regulated by neuronal activity and affects synapse development and function but at later stages than MEF2. (I have modified the canonical, palindromic CREB pattern TGACGTCA in the last 2 bases to allow slightly more permissive matching according to CREB target sequences quoted in the TRANSFAC transcription-factor database [24].) Table **2.1** shows that *CREB* is quite negatively enriched compared to *Mef2*, in particular among the tiled genes (which were chosen as likely targets of MEF2), where MEF2 and CREB would not be expected to function in tandem.

The simple fold enrichment scores in Table **2.1** are fairly reliable quantifications of the motifs' abundance compared to background in the three regions. The *Mef2* variants have high enrichment, especially in the peaks, since that's why MEME rated them highly in the first place, while *Mef2-Perm* is mostly similar to the canonical *Mef2* and *Mef2-Rdm1* is a basically "decommissioned" version of *Mef2*. In the peak regions, YTAWWWWTAR still stands out as unusually enriched—which of course confirms that this is likely the single best target sequence for binding the MEF2 transcription factor.

The magnitudes of the z-scores should perhaps be taken with a grain of salt. Though their trends accord with the fold enrichment values, the very high values in the whole-chr2 and tiled-array gene regions may reflect an inadequacy of Poisson statistics for providing a standard deviation expected from a random background. Or rather, the assumption of randomness is not entirely appropriate to these genomic regions, which after all have nonrandom, highly structured portions characterized by many repeats (in the genome at large) and by promoters and coding regions near genes. Moreover, the low counts of the control motifs in the Mef2 peaks simply indicates that all motifs are not significantly enriched to the extent of the canonical *Mef2* motif.

Of chief interest in the remainder of this analysis will be any association in the Mef2 peaks between enrichment of MEME's *Mef2-HiSco*, which is *a fortiori* high, and any elevation of its typical degree of conservation in this region compared to that of canonical *Mef2* sites (see Sec. **2.4.2**).

2.3 Conservation Scores in MEF2 Peaks

Truly regulatory motif instances are likely to be highly conserved in orthologous genes across multiple genomes, implying that both these genes and these regulatory sequences confer functionality that has been fixed by evolution. Therefore, I assume here once again that conservation can be taken as a proxy for function. In studying the Xie motifs (see Chapter **1**), I used counts of aligned genomes that displayed exact conservation of motif instances at individual sites matching the motif. Here I use instead a single score, from publicly available data, that has been precalculated for each base of the rat reference genome.

2.3.1 PhastCons Scores

Mef2 peaks and motifs are assessed here for their depth of conservation based on published *phastCons Scores* (*PH*ylogenetic *A*nalysis with *S*pace/*T*ime models-derived *Cons*ervation) [25]. Sets of phastCons scores for *rn4* and other genomes are created specifically to be displayed as the "Conservation" tracks on the UCSC Web-based Genome Browser [17].

For a genome of interest, such as *rn4* for rat, the downloadable phastCons scores from Ref. [26] derive from a multiple alignment produced by the program MULTIZ [27]. The published 9-way alignment phastCons9way [26] includes *rn4* (rat, Nov. 2004) plus 8 other vertebrate genomes: *mm8* (mouse, Feb. 2006), *hg18* (human, Mar. 2006), *canFam2* (dog, May 2005), *bosTau2* (cow, Mar. 2005), *monDom4* (opossum, Jan. 2006), *galGal2* (chicken, Feb. 2004), *xenTro1* (frog, Oct. 2004), and *danRer3* (zebrafish, May 2005). Generally, shorter sequences require greater phylogenetic depth and/or more genomes to characterize their conservation [9, 28] accurately. The evolutionary breadth of this set, from human to zebrafish, is comparable to the 17 vertebrate genomes whose alignment was used in Ref. [26]. Thus, these phastCons scores should be a viable guide to conservation of short motifs and hence to MEF2 target functionality in rat.

Each aligned base of the reference genome is evaluated for conservation according to its degree of alignment with the other genomes as calculated by MULTIZ using a phylogenetic hidden Markov model [25]. The calculation models nucleotide substitutions at each site and how these changes compare at neighboring sites according to probabilistically assigned mixtures of states such as "conserved" *vs.* "nonconserved" (i.e., purifying *vs.* neutral substitutions) and "coding" *vs.* "noncoding." Phylogenetic distances between closely related genomes (e.g., rat and mouse) as opposed to evo-

lutionary well separated ones (e.g., rat and zebrafish) are automatically taken into account. MULTIZ tries to align the reference genome with several others that may span a wide evolutionary range. This is a difficult calculation, and not all portions of one genome can be aligned with the others—in part owing to evolutionary divergence, naturally, but also due to the complexity of multiple whole-genome comparisons, incorrect assignments of alignment gaps, false-positive alignments of subsequences, etc. Nevertheless, the phastCons9way scores reported in Ref. [26] cover approximately two-thirds of the bases in the rat genome. Although missing scores might indicate a failure of multiple alignment of orthologs due to genetic novelty, the reasons for unscored genes will not be pursued here.

Bases that can be meaningfully evaluated finally receive scores s whose values lie in the range $s \in [0, 1]$. The distribution of scores across the rat genome is far from uniform or unimodal, however. Figure **2-2(a)** shows the normalized distribution of phastCons over all bases of rat chromosome 2 for which scores were given in Ref. [26] (168,459,582 out of a total 258,207,540 bp of sequence). There is a marked tendency for scores to cluster at the two extremes in a ratio of about 9:1. The genomic average on chr2 is $\bar{s}_{\text{chr2}} = 0.1025$, close to the 10% one would expect were there 9 bases with $s = 0$ for every 1 base with $s = 1$. This behavior and magnitude is replicated on all other chromosomes. Over the whole rat genome I found an average $\bar{s}_{\text{rat}} = 0.0997 \pm 0.0082$, ranging from a minumum average 0.0832 on chr12 to maxima 0.1098 on chr3 and 0.1264 on chrX (see Fig. **2-3**).

Therefore, when a region has an average conservation score of 0.25, say, this can just as well be interpreted as that region having a fraction 0.25 of bases that are well conserved and a fraction 0.75 of bases that are poorly conserved. The same observation holds for conservation scores describing motif occurrences. If a motif such as *Mef2* is represented by the *average* of its 10 bases' individual phastCons scores, it turns out that each motif instance is likely to comprise bases with scores either all $s \to 0$ or $s \to 1$. (Motif instances with one or more unscored bases are ignored.) The average score over multiple motif instances, then, can also be taken to represent the fraction of that motif's instances that are conserved. In Fig. **2-2(b)**, the phastCons scores of all 10-bp windows in the Mef2 peaks (covering 196,264 nt of 199,498 bp of total sequence) still occupy the extremes. In the inset to (b), which has wider histogram bins, the distribution's shape is more clearly seen to be vastly underpopulated in the central 80% basin compared to its edges. (Averages for 11-bp windows in each region are practically the same as for 10-bp windows.)

The only difference from the genome-wide trends is that average phastCons scores for all 10-bp windows in the Mef2 peaks is notably higher at $\bar{s}_{\text{peaks}} = 0.2529$. For comparison, in the tiled-array gene regions (296 genes plus 10kb padding at both ends, 7,361,305 scored bases out of total sequence 31,309,615 bp) the overall average equals $\bar{s}_{\text{genes}} = 0.1480$. The average values $\bar{s}_{\text{rat}}$, $\bar{s}_{\text{genes}}$, and $\bar{s}_{\text{peaks}}$ are summarized in Fig. **2-3**.

Conservation is thus seen to be higher on the average near genes in the genome—

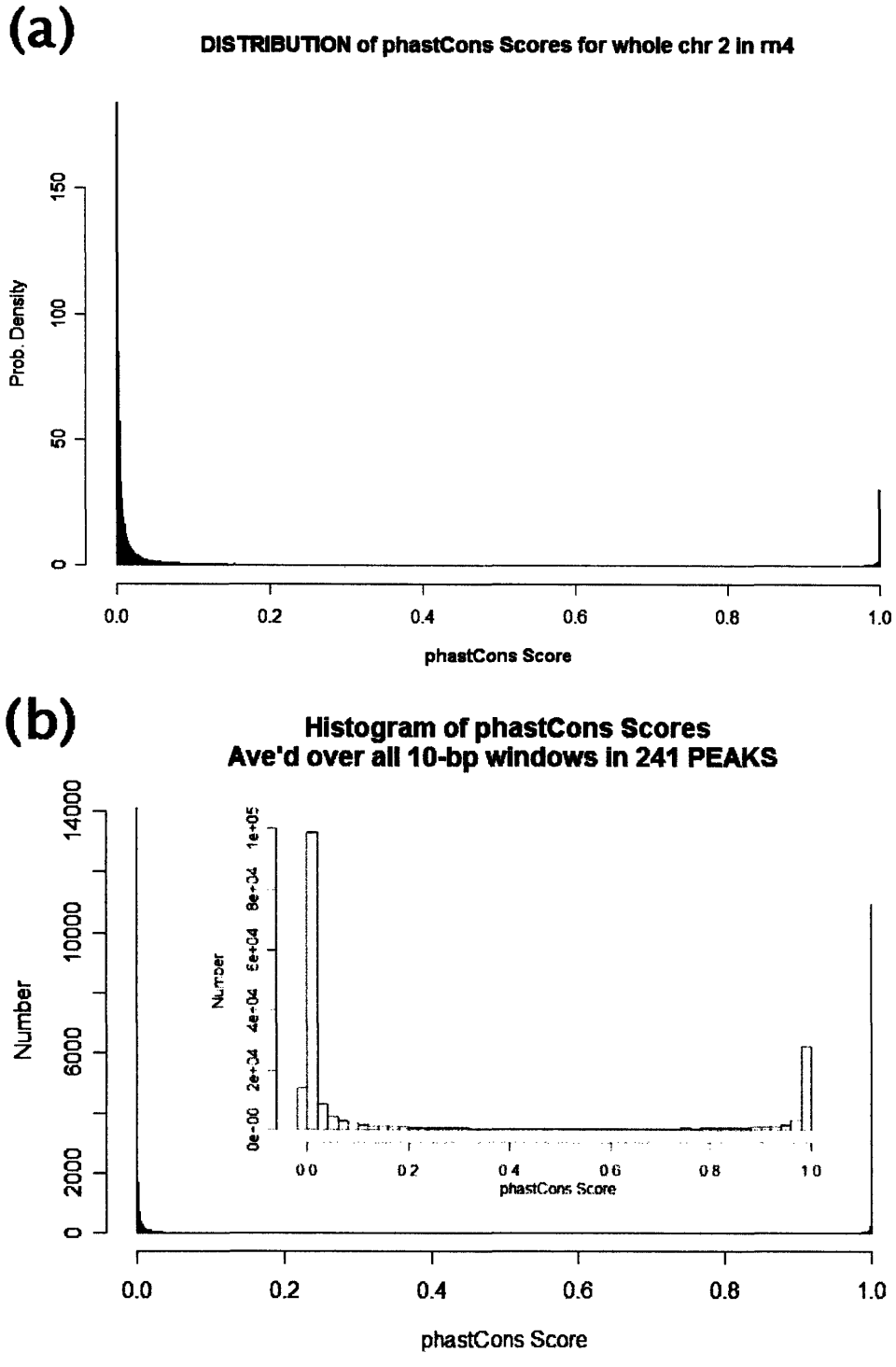

Figure 2-2. Distribution of phastCons scores for **(a)** all of rat chromosome 2, distribution normalized to unity, and **(b)** the Mef2 peaks, histograms with bin size $\Delta s = 0.0001$ and (inset) $\Delta s = 0.02$.

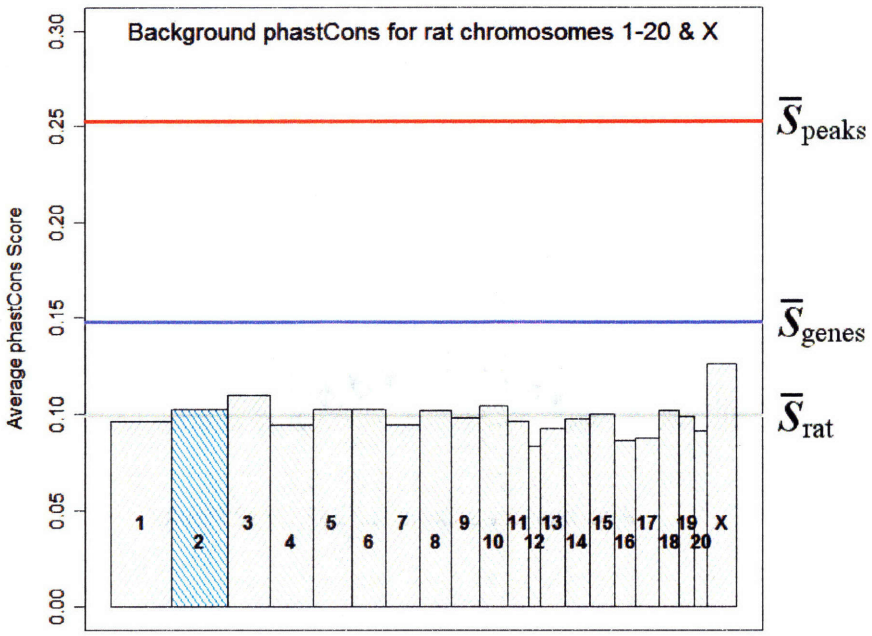

Figure 2-3. Background conservation levels (average phastCons scores): individual rat chromosomes 1–20 and X (histogram width proportional to number of bases scored); $\bar{s}_{\text{rat}} = 0.0997$ (*gray*), genomic weighted average; $\bar{s}_{\text{genes}} = 0.1480$ (*blue*), in tiled array genes (including 10kb padding at 5′ and 3′ ends); and $\bar{s}_{\text{peaks}} = 0.2529$ (*red*), in Mef2 peaks (all 10-bp windows).

unsurprisingly—but is specifically greater in the regions associated with MEF2 binding. This general trend will be borne out even more acutely when considering conservation of instances of *Mef2* and *Mef2*-variant motifs. These *background conservation levels* will serve as controls in their respective regions for studying motif conservation. I will continue to employ chromosome 2 with $\bar{s}_{\text{chr2}} = 0.1025$ (*cyan* histogram in Fig. **2-3**) as a representative background for some genome-wide comparisons.

2.3.2 Conservation Maps in MEF2 Peaks

The 241 Mef2 peaks identified by the ChIP-chip experiments range in length from 336 to 3341 bp. How are regions of higher conservation distributed within this special subset of genomic sequences? Are the more highly conserved bases especially associated with *Mef2* target sequences?

It will prove instructive to visualize the complete set of Mef2 peaks and their conservation properties as a whole. For reference, first consider the mean conservation score over the set of all the tiled-array gene regions, Fig. **2-4(a)**, and over the set of peaks, Fig. **2-4(b)**. Each set has been ordered by score, with histogram widths proportional to bases scored; each rectangle's area thus captures the total number of well conserved bases in that gene

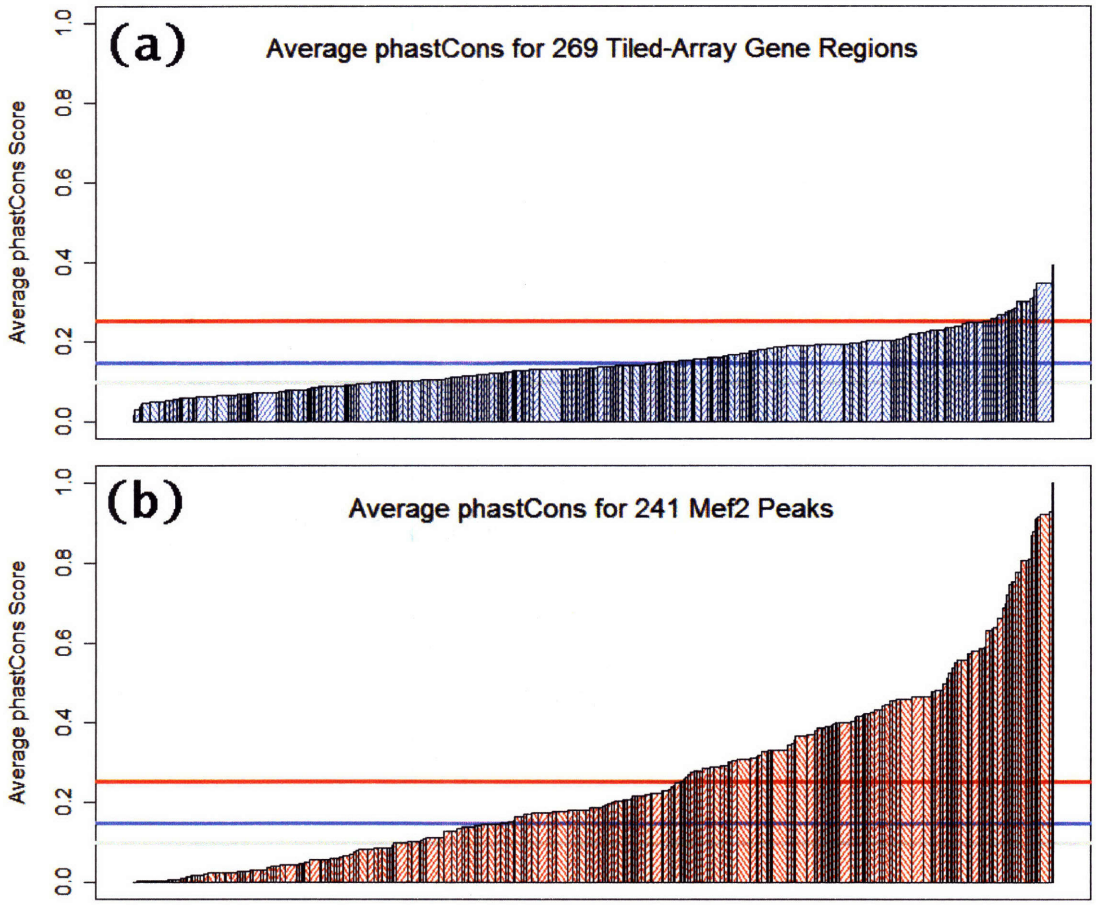

Figure 2-4. Average conservation score for **(a)** each tiled-array gene region with any reported phastCons scores, and **(b)** each Mef2 peak. Histogram widths proportional to bases scored in each gene region or peak. Background conservation levels as in Fig. **2-3**.

region or Mef2 peak. Background levels are marked as in Fig. **2-3**.

Note that a large fraction of the bases in the tiled-array gene regions—`23,948,310` out `31,309,615` bp (`76%`)—have no published phastCons score in `phastCons9way` and so are not included in this figure. In fact, 27 of the 296 tiled-array gene regions have no scored bases at all. Compared to the 123 of the 296 (`41.6%`) tiled genes considered here that are controls, 13 out of these 27 (`48.1%`) scoreless genes were controls. As a random sample this has a one-tailed p-value of `0.298` (Fisher exact test; see Appendix **A.1**), so the distribution of absent phastCons scores is probably not biased with respect to control *vs.* non-control genes.

The lower average $\bar{s}_{\text{genes}} = 0.1480$ over all scored bases in the tiled-array gene regions (*blue*) compared to $\bar{s}_{\text{peaks}} = 0.2529$ for the peaks (*red*) could be accounted for at least in part by regarding the 20 kb of "padding" that was scored in each gene region as typical genomic DNA, with $\bar{s}_{\text{rat}} = 0.0997$ (*gray*). That amounts to 296×20 kb ~ 6 Mb out of the 31 MB of total sequence in the gene regions, or about 20%, which would imply that the

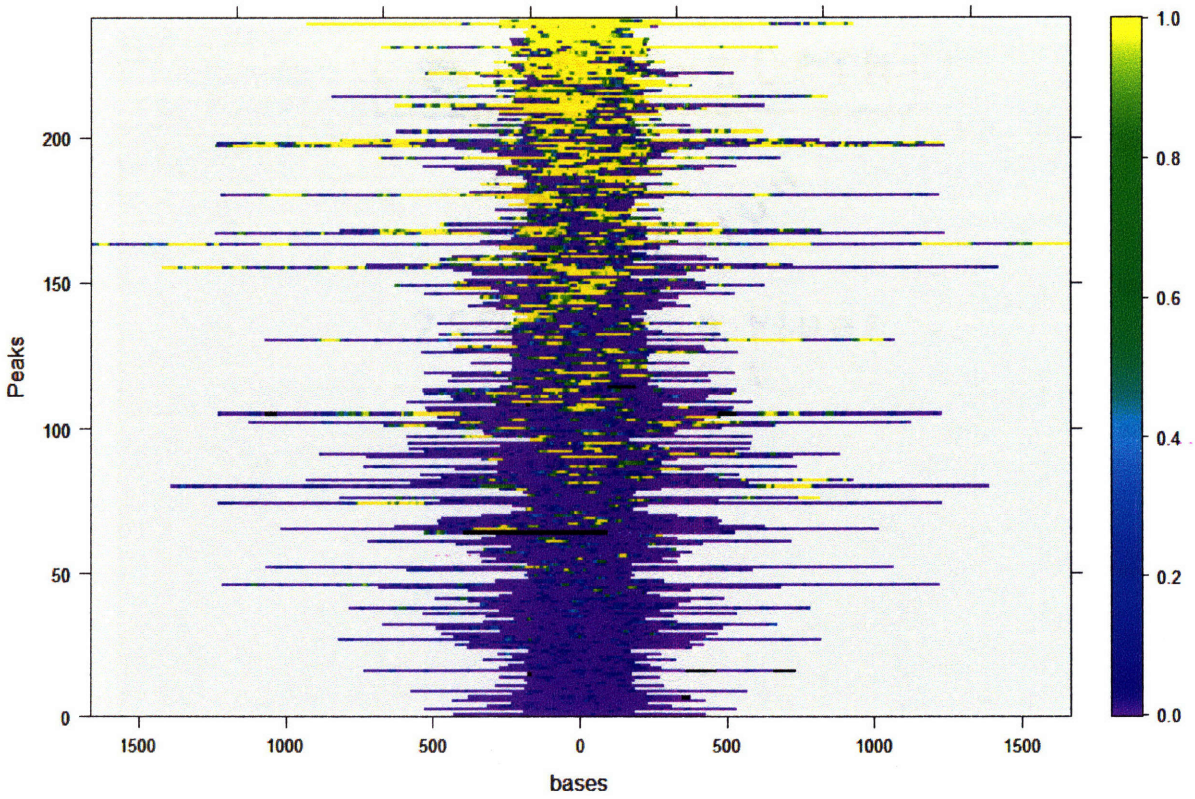

Figure 2-5. Conservation map of phastCons scores for all 241 Mef2 peaks. Sequences are lined up at their centers and ordered left–right from 5′-to-3′ and top-to-bottom from highest to lowest average score. Color key covers full range $s \in [0, 1]$ of possible values; black indicates bases for which no score was available.

remaining 80% has $\bar{s}_{\text{genes/proper}} \approx 0.16$. This is not even as high as $\frac{2}{3}\bar{s}_{\text{peaks}}$, though, so I conclude that *the Mef2 peaks are significantly enriched in highly conserved bases, by at least 50%, over the genes that are themselves targets of MEF2*.

Turning to conservation in the Mef2 peaks in greater detail, phastCons scores for all `199,498` bases of all `241` peaks can be depicted in one go. Such a *conservation map* is shown in Fig. **2-5**. Each horizontal strip represents the DNA sequence of one peak oriented 5′-to-3′ from left to right. Base-by-base scores s are keyed by color from high ($s \to 1$, *yellow*) to low ($s \to 0$, *purple*), with unscored bases blacked out. I've centered the peaks left–right and ordered them from top to bottom according to their average score with highest values placed at top. This is reflected in the highly conserved, mostly yellow peaks in the upper half of the plot and the less conserved, purple peaks in the lower half. (I created this plot and the remaining plots in this Chapter, and performed the related analyses, using the package `R` 2.4.1; see Appendix **B**.) What is immediately clear with regard to conservation is the appearance in this map of uninterrupted segments that are almost entirely either yellow or purple. These peak sequences are, for the most part, partitioned into exclusive

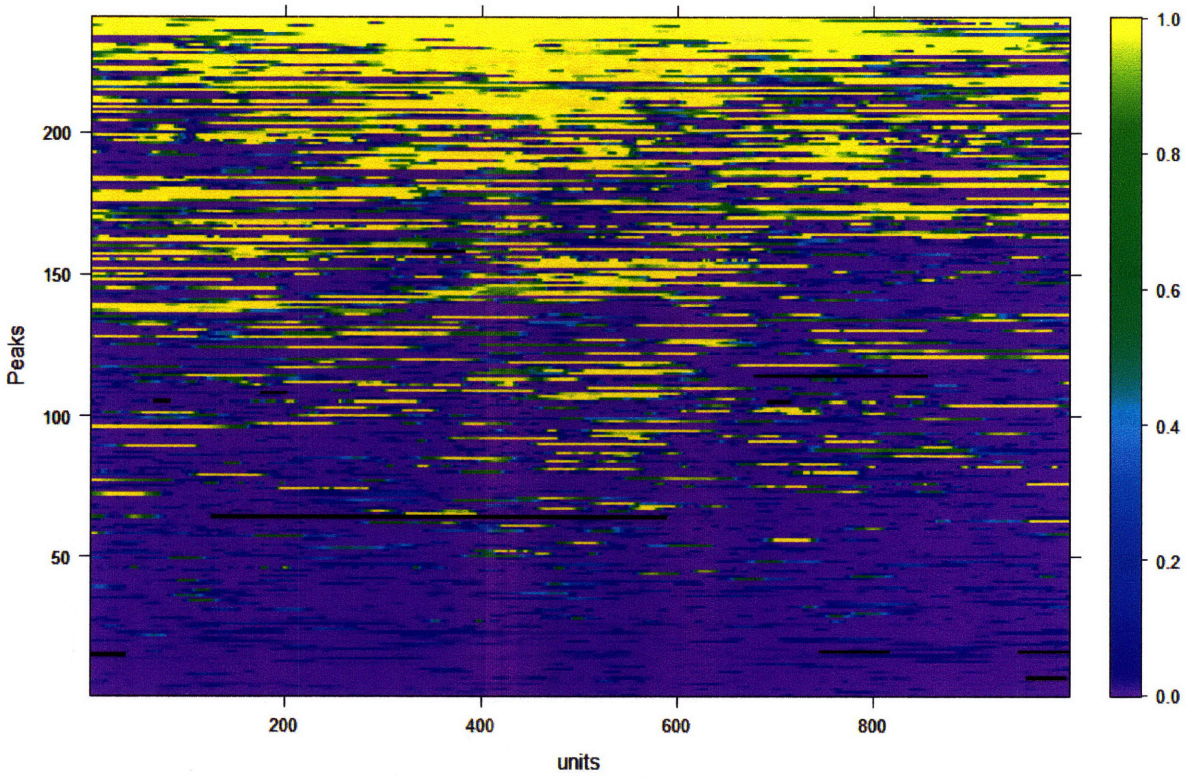

Figure 2-6. Conservation map of phastCons scores for all 241 Mef2 peaks, which have been stretched to cover the same length of 1000 (arbitrary units). Sequences ordered from highest to lowest average score, with conservation color key as in Fig. **2-5**.

islands of high and low conservation. Figure **2-5** provides graphical evidence for the mostly-conserved-*vs.*-mostly-not interpretation of individual bases or motifs in the Mef peaks, and lends support to the interpretation of conservation scores averaged over several motifs as the fraction of motifs that are either well conserved or not.

The Mef2 peaks are expected to associate with loci for functional MEF2 binding, which may in turn be expected to have the greatest frequency near the peak centers. A glance at Fig. **2-5** suggest that this may be so, but islands of high conservation are seen to be scattered away from the centers as well. To compare all the peaks with one another more equitably, their lengths can be "stretched" (or shrunk) to a common, standard length. Such a conservation map of stretched peaks is shown in Fig. **2-6**. The scale used is in arbitary units u over the domain $u \in [0, 1000]$. It is evident that areas of highest conservation are found near peak centers, though not exclusively. There is also no obvious preference for conservation toward $5'$ or $3'$ ends. Conservation scores $\{s(u)\}$ for a peak with N base pairs can be symmetrized about its center via $s_{\rm sym}(u) \equiv \frac{1}{2}\left[s(u) + s(N-u)\right] = s_{\rm sym}(N-u)$. Every stretched peak $\#n$ ($n = 1, \ldots, 241$) can then be regarded as a continuous, symmetric function $s_n^{(\rm str)}(u)$ of scores over the same domain of u. To summarize the conservation

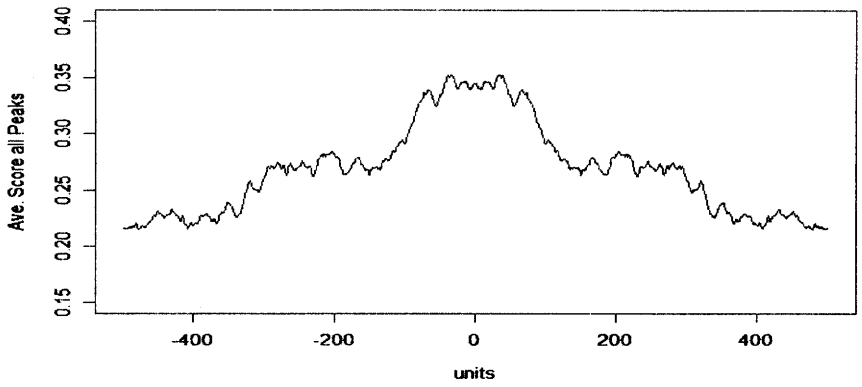

Figure 2-7. Conservation score function $s_{\text{ave}}^{(\text{str})}(u)$ averaged over all 241 symmetrized peaks.

distribution for all peaks, one can simply take the average over all the functions $\{s_n^{(\text{str})}(u)\}$ at each value of u:

$$s_{\text{ave}}^{(\text{str})}(u) = \frac{1}{241} \sum_{n=1}^{241} s_n^{(\text{str})}(u) \tag{2.1}$$

Peaks are weighted equally in Eq. (2.1); they could instead be weighted, e.g., according to their actual lengths in bp. Figure **2-7** shows a plot of $s_{\text{ave}}^{(\text{str})}(u)$, the average stretched-peak distribution of conservation scores. This "typical peak" displays a pronounced bump in conservation within approximately the central 20% of its overall length, at a value of $s_{\text{ave}}^{(\text{str})}(u) \sim 0.35$. This should be compared to the adjacent "plateaus" at $s_{\text{ave}}^{(\text{str})}(u) \sim 0.27$ and the outermost 20% on each end with $s_{\text{ave}}^{(\text{str})}(u) \sim 0.23$, together comparable to the peaks' background level $\bar{s}_{\text{peaks}} \approx 0.25$. Thus, the peaks' central bump is roughly 30--50% higher than the peaks' limbs on average, suggesting that *functional MEF2 binding is likely concentrated near the center of the Mef2 peaks identified by the ChIP-chip experiments of Ref. [1]*, as expected.

It is also instructive to compare the conservation map of Fig. **2-6** with an identical map that additionally depicts specific loci of predicted MEF2 binding. Figure **2-8** displays such a map, with all 151 sites of matches to the canonical *Mef2* motif YTAWWWWTAR indicated as *red* strips. (Each site is 10 bp wide but stretched by a different amount according to the underlying peak's length.) Many of these predicted MEF2-binding loci lie within highly conserved portions of the Mef2 peaks—but a nonnegligible number do not. Conservation properties of these and variant *Mef2* consensus sites are analyzed in the next section.

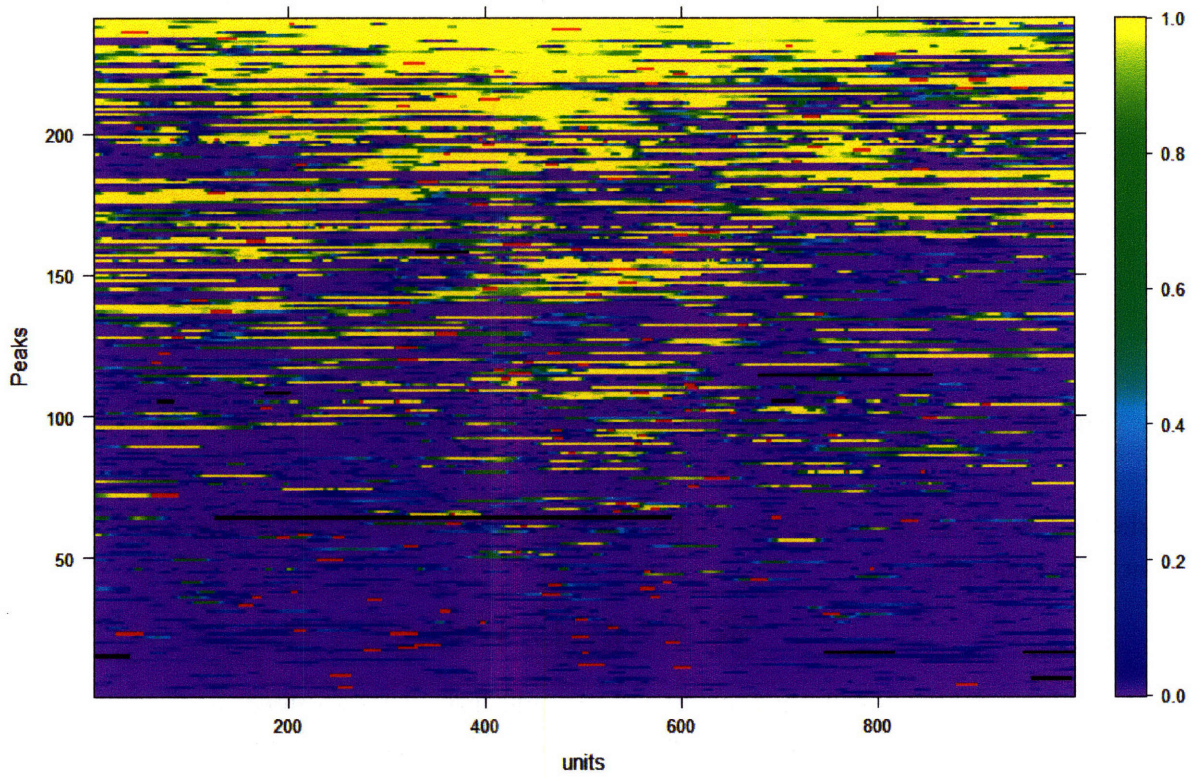

Figure 2-8. Conservation map of phastCons scores for all 241 Mef2 peaks, stretched as in Fig. **2-6**, but including loci that match the canonical *Mef2* consensus binding motif YTAWWWWTAR (*red* strips).

2.4 Conservation Scores near Consensus Sites

Baseline phastCons scores in rat (Fig. **2-3**) were demonstrated in Sec. **2.3.1** to be ordered hierarchically: $\bar{s}_{\text{rat}} \approx \bar{s}_{\text{chr2}} < \bar{s}_{\text{genes}} < \bar{s}_{\text{peaks}}$. It is reasonable to expect that conservation would typically be even greater at loci that match the *Mef2* consensus motif or one of its variants, as many of these sites may be specifically functional. This indeed turns out to be the case. Surprisingly, conservation at these potential binding sites is further associated with a higher average conservation rate in the *neighborhood* of these loci. In this section, I examine conservation properties of the consensus sites themselves against the backgrounds of the rat genome, the tiled-array gene regions, and the Mef2 peaks.

2.4.1 Conservation Distributions near *Mef2* Motifs *vs.* Background

As discussed in Sec. **2.3.1**, the individual bases of any one occurrence of a short motif such as *Mef2* are likely to have similar phastCons scores, whose average $\bar{s}$ is in turn more likely to lie towards either extreme $\bar{s} \approx 0$ or $\bar{s} \approx 1$. A motif's average conservation score in this regard is

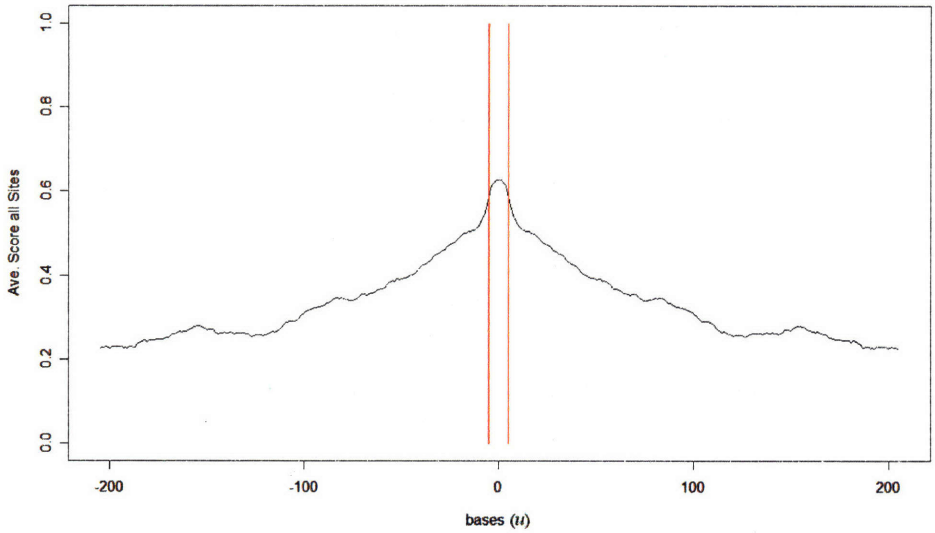

Figure 2-9. Average base-by-base phastCons score for all 151 instances of the canonical *Mef2* motif, and their neighboring sequence, in Mef2 peaks. Red lines mark 10-bp window within which matching sites are lined up. Total 410 bp includes sites' neighboring 200 bp on either side. Score at each base is average over all instances after left–right symmetrization.

reduced to a single number (per genomic region). For example, the 151 instances of *Mef2* in the Mef2 peaks have $\bar{s}_{Mef2} = 0.617$. This is an extraordinarily high average value, indicating that well over 50% of these instances may be functional in the context of actual MEF2 binding. Though perhaps unsurprising, this procedure of reducing the motif to a point site and averaging over many points obscures the fact that the motif operates in the context of its surrounding sequence, probably involving its immediate neighborhood most strongly.

A straightforward way to characterize the motif-matching loci along with their genomic environment is to use their centers as a common reference locus and conduct a base-by-base average over all occurrences out to some suitable distance. That is, instead of focusing one's view on indiscriminantly aligned sequences such as the Mef2 peaks in Figs. **2-5** through **2-8**, one should compare sequences of some size centered on the consensus sites—all the *red* bits in Fig. **2-8**. I illustrate this in Fig. **2-9** for the canonical motif *Mef2*—it should be regarded as a "better" version of Fig. **2-7**. All 151 matches to YTAWWWWTAR in the Mef2 peaks were registered to the 10-bp window delineated by the vertical *red* lines in the figure. On either side of every consensus site, wings 200 bp wide were included to afford a reasonable footprint into the surrounding sequence. (Storing the conservation scores for more than a few hundred bases for every single motif match became impractical and unhelpful for the huge numbers of matches found for the more permissive motives in the large tiled-array gene regions and entire chromosomes.) Scores for each peak instance plus its neighborhood comprise a 410-bp sequence, represented by a function $s_i^{(\text{peaks})}(u)$ $(i = 1, \ldots, 151)$ over the domain $u \in [-205, +205]$. (Each DNA base occupies one unit of u on this scale; the *Mef2*

sites occupy the span $u \in [-5,+5]$.) After each of these functions is left–right symmetrized as before, $s^{(\mathrm{peaks})}_{\mathrm{sym}\,i}(u) \equiv \frac{1}{2}\left[s^{(\mathrm{peaks})}_{i}(u) + s^{(\mathrm{peaks})}_{i}(-u)\right] = s^{(\mathrm{peaks})}_{\mathrm{sym}\,i}(-u)$, they are averaged over all 151 instances at each fixed value of u [cf. Eq. (2.1)]:

$$s^{(\mathrm{peaks})}_{\mathrm{ave}}(u) = \frac{1}{151}\sum_{i=1}^{151} s^{(\mathrm{peaks})}_{\mathrm{sym}\,i}(u) \tag{2.2}$$

As noted above, this average score over the consensus sites, $u = [-5,+5]$, attains the value $\bar{s}_{Mef2} = 0.617$, which marks the prominent central "nose" in the center of Fig. **2-9**. This result elaborates on the noted prevalence for many of the *red Mef2* segments in Fig. **2-8** to lie within the highly conserved *yellow* branches of the Mef2 peaks. Remarkably, the maximum score at the consensus is accompanied, on the average, by about 100 bp of sequence on either side that rises up from the peak background level $\bar{s}_{\mathrm{peaks}} \approx 0.25$. In other words, *in MEF2 binding peaks a rather high likelihood of conservation at Mef2 consensus sites is associated, on average, with a significant degree of conservation within at least 50 bp on either side of these sites.* The fact that conservation rates are elevated within a 100–200-bp neighborhood of sites that probably have some function related to MEF2 binding suggests that other factors may play a role in the binding of MEF2, including the possible binding of those factors near these loci.

To confirm the significance of conservation near *Mef2* consensus sites in the Mef2 peaks, the average phastCons scores I've calculated for Fig. **2-9** need to be compared to the various controls introduced in Sec. **2.3.1**. It is of particular interest to check whether *Mef2* instances are highly conserved in genes in general and in the rat genome at large—i.e., in the tiled-array gene regions pertinent to this study, and in representative genomic sequence not specific to the ChIP-chip experiments, e.g., rat chromosome 2 or some random portion thereof. To this end, I extracted all 7443 instances (Table **2.1**) of the canonical *Mef2* motif plus their neighboring ±200 bp from the 31 Mb of sequence in the 296 tiled-array gene regions, a search that extended up to 10 kb beyond both ends of these genes. (The `perl` and `R` scripts I wrote are summarized in Appendix **B**.) The approximately 5800 instances that remained after discarding instances that contained any unscored bases were symmetrized and averaged, just as per the peaks' instances, to produce an instance-neighborhood average scoring function $s^{(\mathrm{genes})}_{\mathrm{ave}}(u)$ for the gene regions analogous to $s^{(\mathrm{peaks})}_{\mathrm{ave}}(u)$ for the peaks.

For the genomic controls, the number of instances of *Mef2* on chromosome 2 was larger than in the gene regions by an order of magnitude, $\sim$ 70,000 of them in $\sim$ 260 MB; for the more permissive variant motifs this number was yet another order larger (see Table **2.1**). Noting that storage of 410 individual phastCons scores for, e.g., the approximately 900,000 instances of *Mef2-HiSco* on chr2 would nominally require over 1 GB of disk storage, I decided that probing the whole chromosome would be overkill and that a subset of chr2 would suffice for purposes of sampling each motif. (Sampling average scores from different

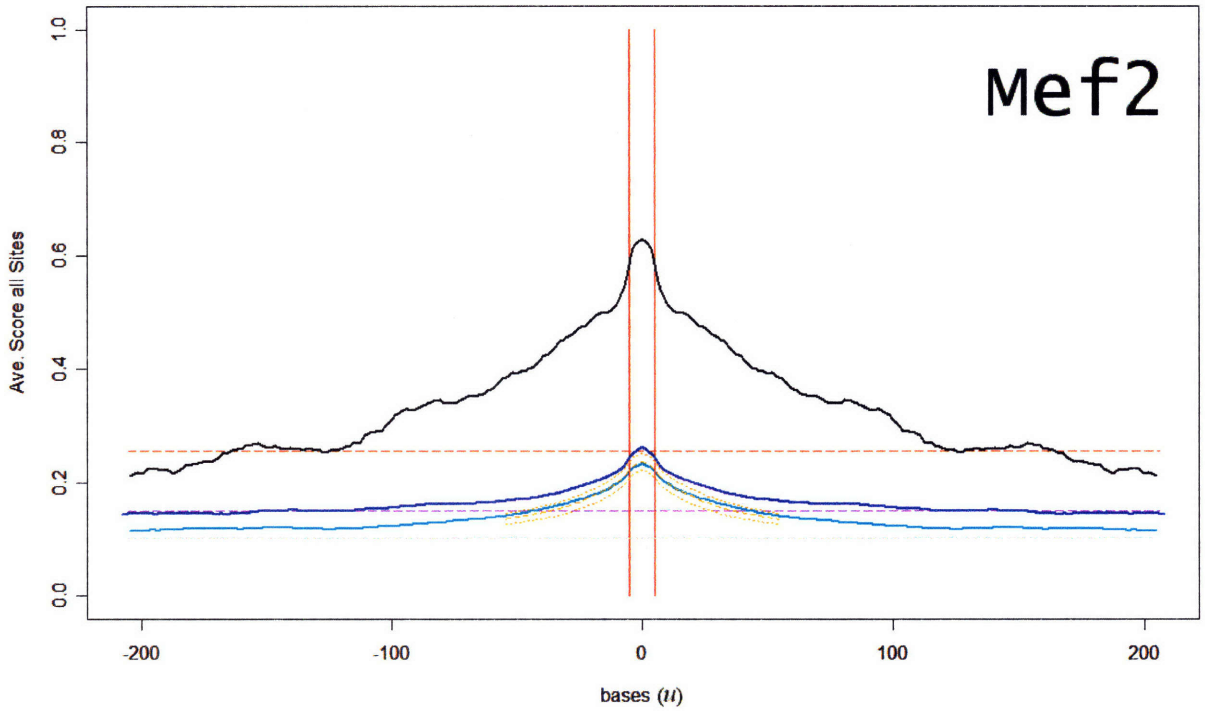

Figure 2-10. Average base-by-base phastCons scores for all instances of the canonical *Mef2* motif YTAWWWWTAR, and their neighboring sequence, as in Fig. **2-9**: over 10% of rat chromosome 2, $s_{\text{ave}}^{(\text{chr2})}(u)$ (*cyan* curve); in tiled-array gene regions, $s_{\text{ave}}^{(\text{genes})}(u)$ (*blue* curve); in Mef2 peaks, $s_{\text{ave}}^{(\text{peaks})}(u)$ (*black* curve). Average background scores, dashed horizontal lines: $\bar{s}_{\text{chr2}}$ (*gray*); $\bar{s}_{\text{genes}}$ (*purple*); $\bar{s}_{\text{peaks}}$ (*red*). Conservation function $s_{\text{ave}}^{(\text{rat})}(u)$ within 50 bp of all instances, mean $\pm$ 1 standard deviation (dotted *gold* curves) covering all chromosomes.

genomic regions did not change any of the present results significantly.) Thus, 10% of rat chr2 was found to have about 6800 instances of *Mef2*, of which about 4000 had no unscored bases. I extracted scores for these plus their 400-bp neighborhoods, which were then symmetrized and averaged to form the function $s_{\text{ave}}^{(\text{chr2})}(u)$ representing unspecific rat genome conservation near *Mef2* sites.

The average conservation functions $s_{\text{ave}}^{(\text{peaks})}(u)$, $s_{\text{ave}}^{(\text{genes})}(u)$, and $s_{\text{ave}}^{(\text{chr2})}(u)$ are shown in Fig. **2-10** over their 410-bp domain of u, centered on the 10-bp consensus sites as in Fig. **2-9**. These curves can be interpreted as the fraction of matches to the canonical MEF2 consensus YTAWWWWTAR that are well conserved in the Mef2 peaks, in the tiled-array gene regions, and in the whole rat genome, respectively. The horizontal dashed lines mark the control values for each region ($\bar{s}_{\text{peaks}}$, $\bar{s}_{\text{genes}}$, and $\bar{s}_{\text{chr2}}$) discussed in Sec. **2.3.1**. The *gold* curves in the central 100 bp show the mean and spread of genomic conservation scores $s_{\text{ave}}^{(\text{rat})}(u)$ near *Mef2* sites over *all* rat chromosomes, indicating that the chr2 curve is typical of the rest of the genome.

Figure **2-10** *encapsulates the central results of this study.* All three conservation curves rise above their respective random-background values—but by far the effect is widest and highest in the Mef2 peaks (same curve, *black*, as in Fig. **2-9**). The gene-region curve (*blue*) is uniformly $\Delta s \sim 0.03$ higher in value than the one for chr2 (*cyan*), but they seem otherwise equivalent. Conservation is, as one would expect, better when genes are specifically included in the mix, though not much here because the gene regions consist mostly of noncoding sequence bracketing the genes. The "conserved-by-association" effect of the *Mef2* consensus sites on their surrounding sequence is also not nearly as dramatic in the gene regions and across chr2 as in the Mef2 peaks, but it is not gone entirely. At the sites themselves (i.e., between the red lines), the average scores equal 0.255 and 0.228 in these regions, respectively, *vs.* 0.617 in the peaks. For chr2 this is about double the background value $\bar{s}_{\text{chr2}} = 0.100$. Though confined to within only 10–20 bp of the sites, it is interesting that *conservation of Mef2 motifs stands out even in the genome at large*. The large majority of the many consensus binding sites for MEF2 presumably are not functional loci of MEF2-controlled transcription, so it is not entirely clear why selective evolutionary pressure would maintain these sequences. There is some experimental evidence that MEF2 does indeed bind chromatin in the promoters of a large number of genes irrespective of whether neurons are stimulated [1, 29], possibly playing a role in the recruitment of RNA polymerase II; the MEF2 remains in an inactive state until dephosphorylated in an activity-dependent manner, which then leads to initiation of transcription. My observation of a small but significant, nonrandom preference for the generic conservation of *Mef2* motifs is consistent with a biological interpretation that widespread MEF2 binding throughout the genome may function in recruiting the transcriptional machinery to gene regions.

2.4.2 Alternative Motifs—Conservation near *Mef2* Variants

The exceptional conservation rates found for the canonical *Mef2* motif in the rat genome raises the question of whether other highly enriched motifs might also have high phastCons scores and, by implication, likely functionality in Mef2 peaks and in other genomic regions. The most enriched motif predicted by MEME in the Mef2 peaks, *Mef2-HiSco*, is listed in Table **2.1** as having about the same 1.3-fold enrichment as *Mef2* in the tiled-array gene regions and across chr2 and—by construction—the highest fold-enrichment of any motif in the peaks besides *Mef2* itself. It is also by far the most permissive motif analyzed here, found on average about every 1 in 300 bases on chr2. I calculated average phastCons scores for *Mef2-HiSco* in the three regions of interest as described above for *Mef2*; the results, analogous to Fig. **2-10**, are displayed in Fig. **2-11**.

One notices a moderate enhancement of these instances, with an the average value $\bar{s}_{\textit{Mef2-HiSco}} = 0.428$ in Mef2 peaks (*black* curve). This clearly lies significantly above the background value $\bar{s}_{\text{peaks}} = 0.2529$, but is less acute and has less apparent influence on the

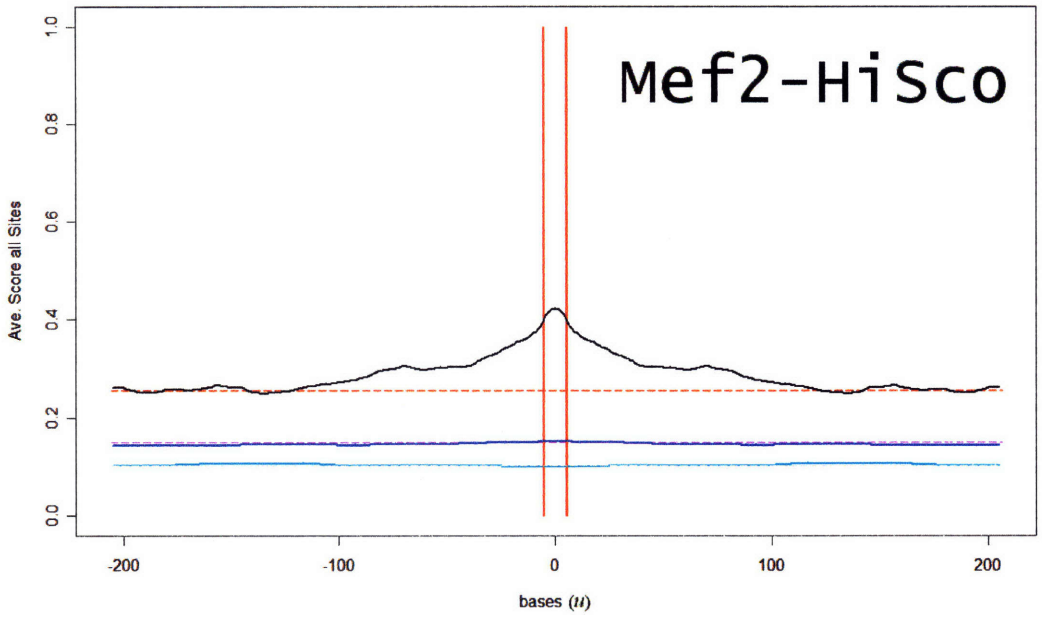

Figure 2-11. Average base-by-base phastCons scores for all instances of the *Mef2-HiSco* motif BTWTWTHWDDH. Curves and background levels colored the same as in Fig. **2-10** for chr2, tiled-array gene regions, and Mef2 peaks.

instances' surrounding neighborhood than *Mef2*. As for *Mef2*, there seems to be a moderate conservation "shelf" in the region roughly 50–100 bp on either side of the consensus sites, and only statistical variations from the background beyond that. But any "neighborhood effect" is *absent* for the tiled-array genes and the genomic background (*blue* and *cyan* curves)—both are spot on their background values $\bar{s}_{\text{genes}}$ and $\bar{s}_{\text{chr2}}$. It is likely that the conservation bump for *Mef2-HiSco* in the Mef2 peaks stems from those matches to BTWTWTHWDDH that are highly similar to YTAWWWWTAR (an analysis to test this was not performed here, however). Thus, the role of transcriptional potentiation that might be attributed to MEF2 binding of DNA does not not appear to be attributable to the variant motif *Mef2-HiSco*, at least not outside the explicit binding regions of the Mef2 peaks.

I have likewise analyzed the phastCons scores in rat for all instances of the other *Mef2*-variant motifs and controls listed in Table **2.1**. Figure **2-12(a)–(f)** shows these motifs' average conservation scores across chromosome 2, in the tiled-array gene regions, and in the Mef2 peaks as in the previous two figures. Recall that *Mef2-var1* was returned by MEME given Mef2 peaks that already contained a match to the canonical *Mef2* consensus, while *Mef2-var2* and *Mef2-var3* was returned for peaks that had no match to *Mef2*, and *Mef2-HiSco* was returned with all peaks as input. The curves in **(a)–(b)** appear to simply recapitulate those for *Mef2-HiSco* (Fig. **2-11**), whereas **2-12(c)** is more suggestive of *Mef2* itself (Fig. **2-10**). The much more common motifs *Mef2-var1* and *Mef2-var2* may simply

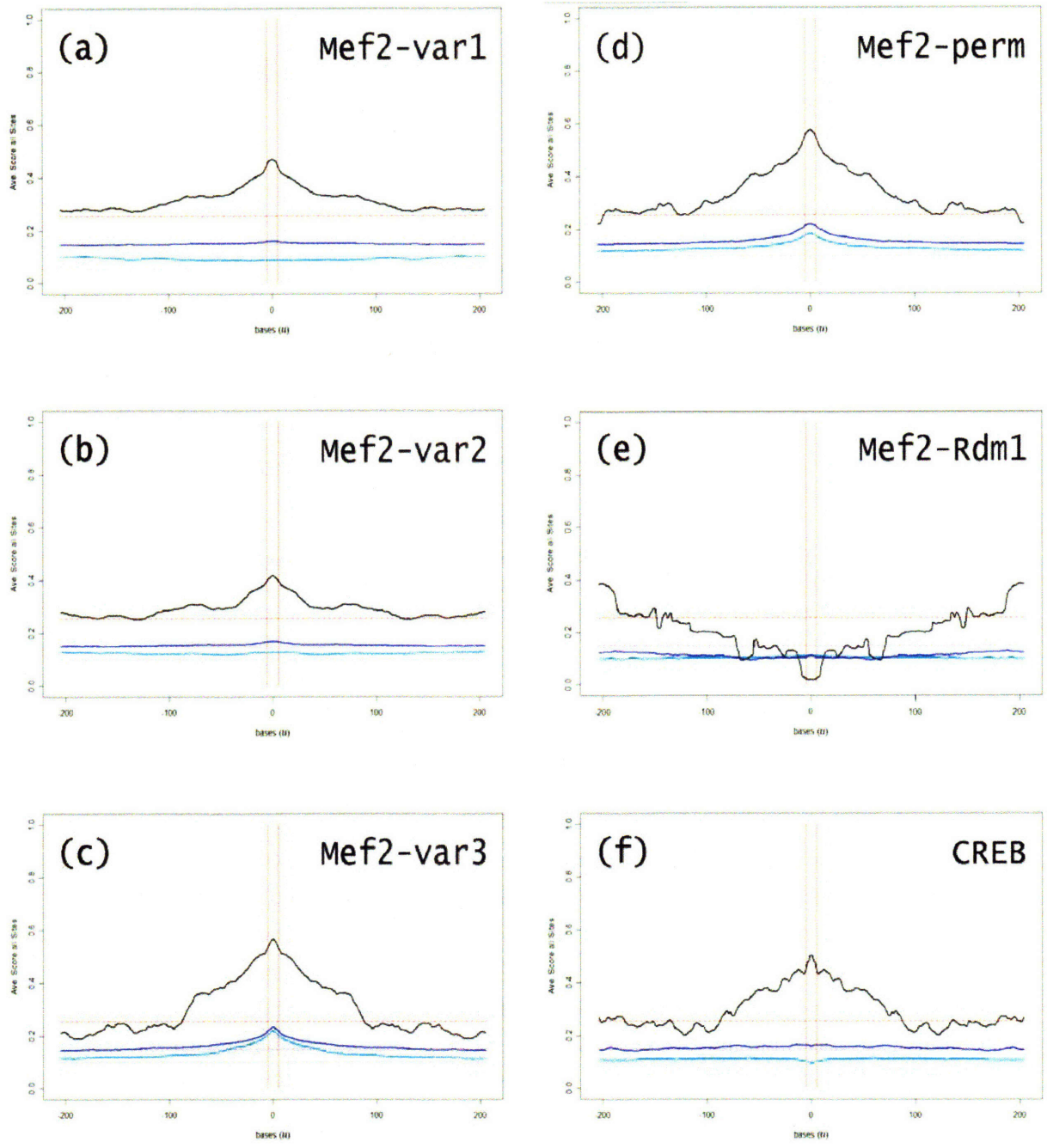

Figure 2-12. Average base-by-base phastCons scores for all instances of the variant and control motifs: **(a)** *Mef2-var1* (HWAWAWWWAR), **(b)** *Mef2-var2* (KBYTDTTTWDD), **(c)** *Mef2-var3* (DRTWWTTWTAR), **(d)** *Mef2-Perm* (WYWAATRWTW), **(e)** *Mef2-Rdm1* (AKCTWWAGMT), **(f)** *CREB* (TGACGTMD). Curves and background levels colored the same as in Figs. **2-10** and **2-11**.

capture the widespread behavior of *Mef2-HiSco*, i.e., moderately elevated conservation in all the Mef2 peak regions but nothing significant otherwise. *Mef2-var3* is in fact the rarest variant enshrined by MEME as "enriched" here, yet it most closely reflects the significant enhancement of conservation near consensus sites seen for the canonical consensus *Mef2*. It is tempting, though perhaps vague, to assign to *Mef2-var3* whatever features and functions of *Mef2* are not accounted for by *Mef2-var1* and *Mef2-var2*. (But note that greater fluctuations in the curve for Mef2 peaks is to be expected for *Mef2-var3* near background levels owing to the low number of *Mef2-var3* counts.) It is not obvious how this might be explained directly from the consensus patterns for these motifs, however.

The other three motifs present interesting differences. In **2-12(d)**, the *Mef2*-permuted motif *Mef2-Perm* looks very nearly the same as Fig. **2-10** for conservation of *Mef2* itself, though fewer counts introduce greater variability, as for *Mef2-var3*. This suggests that it may be the base composition of the canonical *Mef2* consensus that lends MEF2 binding sites their specificity and not merely the order of the bases. There is some evidence for this, which is explored thoroughly in the next section, though the particular pattern YTAWWWWTAR will turn out to be the "best" in the context of the present data. In **2-12(e)**, the (quasi)random motif *Mef2-Rdm1* basically reflects background levels. All its jitter in Mef2 peaks falls within the noise expected for the mere 5 instances there (Table **2.1**). The depression of the scores in the tiled-array gene regions towards the level of the whole-genome background is perhaps explained by the fact that this motif is negatively enriched in the genome; its 30% CG content falls well below the background level of $p_{CG} \approx 0.42$.

Finally, the *CREB* conservation results in Fig. **2-12(f)** are unusual. (Again note that high variability of scores in the Mef2 peaks stems from low counts.) In Mef peaks its average score is enhanced near its matching sites and, like the *Mef2* motif, falls to background levels outside about ±50 bp. The dip in conservation at genomic sites could be statistical in nature; in any case the behavior of its average scores seem to be orthogonal to the substantial de-enrichment of *CREB* consensus sites in the genes and in the whole genome (Table **2.1**). *Nevertheless, both CREB instance enrichment and conservation are distinctly higher in regions associated with MEF2 binding.* As noted above, CREB and MEF2 both influence synapses in an activity-dependent manner. Both factors are expressed in a wide variety of cell types. Whether the association I've described arises from any direct or indirect interaction between these factors remains to be studied.

The statistical significance of the high average conservation score for canonical *Mef2* instances in the Mef2 peaks can be quantified if we once again assume that each instance is either conserved ($\bar{s} \to 1$) or not ($\bar{s} \to 0$). Then the average $\bar{s}_{Mef2} = 0.617$ implies that, of the 151 occurrences of *Mef2*, a total of $N_{\text{cons}} = 0.617 \times 151 = 93.2$ are specifically conserved. With a background conservation score $p = \bar{s}_{\text{peaks}} = 0.253$, however, out of a random sample

Score range:	[0-0.02]	(0.02-0.20]	(0.20-0.80)	[0.80-0.98)	[0.98-1]	TOTAL
Mef2	26	20	25	19	61	151
10-bp Bkgd.	112890	25581	18126	11779	27888	196264
Mef2-HiSco	189	107	83	57	148	584
11-bp Bkgd.	112307	25796	18463	11852	27597	196015

Table 2.2. Distributions of phastCons scores s for instances of *Mef2* (YTAWWWWTAR) and *Mef2-HiSco* (BTWTWTHWDDH) in Mef2 peaks, as well as for background distributions from all 10-bp and 11-bp windows. See text for results of χ^2 test. [*N.B.* Some overlapping instances are included for *Mef2-HiSco*, hence its total differs from Table **2.1**.]

of $N = 151$ we would expect to find only $N_{\text{expect}} = pN = 0.253 \times 151 = 38.2$ of them conserved. Since this is a binomial distribution (Appendix **A.1**), it is well approximated by a gaussian with one standard deviation given by $\sigma = \sqrt{p(1-p)\,N} = \sqrt{0.747 \times 38.2} = 5.3$. The z-score (deviation from the expected mean in units of standard deviations) for *Mef2* motifs in Mef peaks thus equals

$$z_{Mef2} = \frac{N_{\text{cons}} - N_{\text{expect}}}{\sigma_{\text{expect}}} = \frac{93.2 - 38.2}{5.3} = +10.3 \tag{2.3}$$

which implies a p-value $= 3.5 \times 10^{-25}$ ... which is plenty significant.

For comparison with the permissive *Mef2-HiSco* motif predicted by MEME, we can similarly calculate a z-score for instances of *Mef2-HiSco* in Mef2 peaks. Here there are $N = 549$ occurrences, an average score $\bar{s}_{Mef2\text{-}HiSco} = 0.428$ for these instances, and the same background $p = \bar{s}_{\text{peaks}}$. We find $N_{\text{cons}} = 0.428 \times 549 = 235.0$, $N_{\text{expect}} = pN = 138.8$, $\sigma = \sqrt{0.747 \times 138.8} = 10.2$, so

$$z_{Mef2\text{-}HiSco} = \frac{N_{\text{cons}} - N_{\text{expect}}}{\sigma_{\text{expect}}} = \frac{235.0 - 138.8}{10.2} = +9.4 \tag{2.4}$$

and a p-value $= 2.7 \times 10^{-21}$, also highly statistically significant.

A more accurate assessment than a binomial conserved/NOTconserved statistic should compare the actual distribution of phastCons scores for these *Mef2* or *Mef2-HiSco* instances in the peaks to that of the background [Fig. **2-2(b)**]. I binned the scores into five ranges chosen to capture extreme, not-as-extreme, and intermediate values, producing the counts shown in Table **2.2**. Using a χ^2 test ($\nu = 4$ degrees of freedom) yields $p = 3.4 \times 10^{-28}$ ($\chi^2 = 135.0$) for *Mef2* and $p = 1.9 \times 10^{-32}$ ($\chi^2 = 154.8$) for *Mef2-HiSco*. Occurrences of these motifs in MEF2-binding peaks are therefore not only indisputably biased towards high conservation, but they are moreover depopulated of low phastCons scores in favor of intermediate scores.

2.4.3 Conservation and Enrichment of Motif Permutations

In Fig. **2-12(d)** the single permutation WYWAATRWTW of the canonical MEF2 consensus sequence YTAWWWWTAR was seen to echo the principal conservation features of the unpermuted *Mef2* instances plotted in Fig. **2-10**. It is only enriched about 2-fold *vs.* 10-fold for *Mef2*, but its 55 instances in Mef2 Peaks have an average phastCons score $\bar{s} \approx 0.55$—almost as high as $\bar{s}_{Mef2} = 0.62$ itself. More strikingly, it also displays significant conservation rates beyond background around its occurrences in gene regions and throughout the rat genome.

This permutation was selected randomly. Was this just a lucky pick, or does the particular set of *Mef2* base possibilities $\{[A]^2, [A,G]^1, [T]^2, [A,T]^4, [C,T]^1\}$ itself have an inherently high level of conservation for some reason? Of course there is empirical evidence [20, 21] pointing to the particular palindromic motif YTAWWWWTAR as a binding site for MEF2. There is no obvious biological reason for a transcription factor to abandon sequence specificity in this way. But in light of this one case, it would be interesting to at least characterize conservation of alternatives to the *Mef2* motif that have the same base content in different order. Along with conservation scores I will survey the enrichment of *all* permutations of *Mef2*, keeping track of how changes in these quantities depend on sequence similarity to the original consensus. For comparison, I will also survey conservation, enrichment, and similarity among all permutations of the more permissive 11-bp *Mef2-HiSco* motif BTWTWTHWDDH in the Mef2 peaks.

For the canonical *Mef2* motif, the number of distinguishable permutations of the 10 symbols $\boxed{A}\boxed{A}\boxed{R}\boxed{T}\boxed{T}\boxed{W}\boxed{W}\boxed{W}\boxed{W}\boxed{Y}$ equals a multinomial coefficient, $\frac{10!}{2!1!2!4!1!} = 37{,}800$. I generated a list of all 37,800 permutations, and automated the task of performing the following steps for every permuted motif m:

- calculate the expected number of occurrences of the motif in random sequence based on the CG content in Mef2 peaks;
- find all instances of the motif in the Mef2 peaks;
- calculate z_m, an enrichment z-score (Appendix **A.1**), from the expected and found numbers of instances;
- from the found instances' genomic loci, look up the phastCons scores for each instance's 10 bases and calculate an average score s over all instances;
- calculate a similarity score λ_m, discussed below, for the permuted motif sequence *vs.* the unpermuted sequence YTAWWWWTAR.

The similarity score $\lambda(S_1, S_2)$ between two sequences S_1 and S_2 should be a measure of how well their base patterns correlate with one another. The method I use here is based on an algorithm outlined in the supplementary material of Ref. [4]; details are described

here in Appendix **A.3**. Scores lie in the interval $\lambda \in [0,1]$, with only an identical match having $\lambda = 1$ for maximum similarity. A flat background ($p_{CG} = \frac{1}{2}$) would have $\lambda = 0$. The background corresponding to the Mef2 peaks, with $p_{CG} = 0.489$, has $\lambda = 0.0192$ when compared to the *Mef2* motif (or to any of its permutations).

Figure **2-13(a)** plots the results for each permutation as a point (s_m, z_m) with enrichment z-score on the vertical axis and phastCons score s, averaged over all instances, on the horizontal axis. The mean values of z and s over all 37,800 permuted motifs are shown as dashed *gold* lines, with ± 1 standard deviations shown as dotted *gold* lines. The background conservation level for 10-bp windows, $\bar{s}_{\text{peaks}} = 0.2529$ is marked by a gray line. Similarity scores for each permuted motif *vs.* the unpermuted YTAWWWWTAR is represented by the color of the dot, out of a spectrum of colors from *purple* ($\lambda = 0$) to *red* ($\lambda = 1$).

It is immediately apparent from the figure that enrichment and conservation scores for the bulk of the permuted motifs each fall within a quasi-normal distribution. Conservation-s is clustered around $\bar{s} \pm \sigma_s = 0.438 \pm 0.074$, enrichment-$z$ around $\bar{z} \pm \sigma_z = 4.05 \pm 3.30$ (while fold-enrichment *per se* averages 1.77 ± 0.63). The coefficient of determination for all points $\{(s_m, z_m),\ m = 1, \ldots, 37{,}800\}$ is $r^2 = 0.0219 \pm 0.0015$ (squared Pearson-r correlation), indicating that enrichment and conservation have a minor though statistically significant dependence. The "cloud" of points above $z = \bar{z} + 2\sigma_z \approx 11$ does amount to about 5% of all points, as it would for a normal distribution; it is not mirrored below $\bar{z}$ only because lower negative z-scores are truncated by the absence of enrichment less than 0-fold.

On the other hand, it is notable that almost all permutations have conservation values significantly above the background. Moreover, 62.8% of the permutations have greater than 1.5-fold raw enrichment above background. So this whole set of permutated motifs seems to capture base patterns that are both unusually prevalent and evolutionarily maintained, at least in Mef2 peaks. The unpermuted cananonical *Mef2* motif itself (circled point) is among the highest-conservation points. And YTAWWWWTAR is a true outlier on the enrichment axis—*by far the single most enriched motif of all the permuted motifs*. In this regard, the *Mef2* motif sequence stands out as unique among all of its permutated sequences for its highly significant enrichment *and* conservation. The vast majority of its permutations would not be expected to function as transcription factor binding sites, however. Any other biological role for this set of permuted motifs remains an open question.

Do variations in enrichment and conservation depend on how similar the permuted motifs are to the original sequence YTAWWWWTAR? A careful examination of Fig. **2-13(a)** reveals that the *yellow*-through-*red* points, with similarity scores approaching $\lambda \to 1$, appear to be "dragged" to the right and up towards the reference point itself. Overall correlation of similarity scores with conservation s-scores is $r_s = 0.2287 \pm 0.0050$ (Pearson-r) for the whole set; correlation of similarity scores with enrichment z-scores is $r_z = 0.2116 \pm 0.0050$. Although both of these are still low ($r^2 \sim 0.05$), they are significant and do hint at

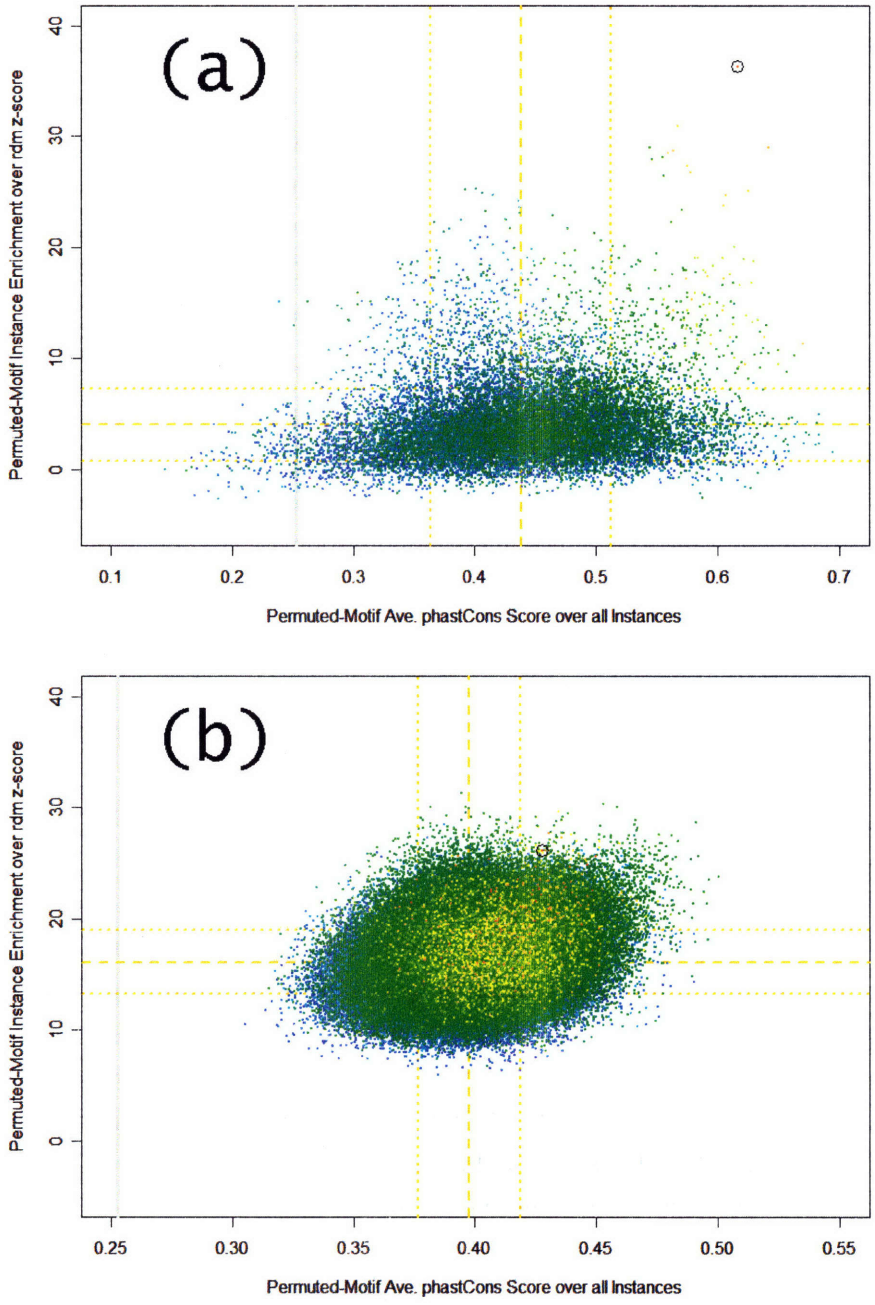

Figure 2-13. Enrichment *vs.* conservation in Mef2 peaks for **(a)** all 37,800 permutations of the *Mef2* motif YTAWWWWTAR and **(b)** all 277,200 permutations of the *Mef2-HiSco* motif BTWTWTHWDDH. One dot per permutation represents the z-score for enrichment over background and phastCons score averaged over all instances in Mef2 peaks; mean scores over all permutations, ± 1 SD, dashed *gold* lines. Background phastCons score, *grey* line. Similarity to unpermuted motif, score λ (defined in text) rainbow-coded from *red* for perfect similarity ($\lambda = 1$) to *purple* for background value ($\lambda \approx 0.02$). Circled point, unpermuted *Mef2* or *Mef2-HiSco* motif.

slightly higher conservation and enrichment the greater a motif's similarity to the canonical YTAWWWWTAR. When I repeat these correlation calculations but limit the motifs to those having the highest 10% of values for s or z, respectively, I find $r_s = 0.1010 \pm 0.0162$ and $r_z = 0.2625 \pm 0.0162$. Thus, the most enriched permutations in Mef2 peaks tend to be most similar to the canonical consensus, while there is less correlation between high conservation and canonical similarity. If anything, the particular order of the bases seems to be more predictive of how often the permuted motif turns up in Mef2 peaks, while conservation of their occurrences is less tied to base order than to base composition.

Turning to the *Mef2-HiSco* motif BTWTWTHWDDH identified by MEME as superbly enriched in Mef2 peaks, we now have to contend with a much more permissive and easily scrambled 11-bp sequence. The set of bases B|D|D|H|H|T|T|T|W|W|W has $\frac{11!}{1!2!2!3!3!} = 277{,}200$ different permutations. Note that calculation and storage of so many instance files, while not onerous, might merit a more sophisticated design for another motif with, say, millions of permutations and/or a larger genomic region containing many more motif instances.

In Fig. **2-13(b)** I again plot each permutation's conservation and enrichment scores (s_m, z_m) as a color-coded point to show the similiarity score λ of permuted motif m *vs.* BTWTWTHWDDH. (There is a high density of points in the center of the "ball"; layers of low-similarity, *purple*, *blue*, *cyan*, etc., points were plotted first and so are obscured by high-similarity upper layers of *green* and *yellow*.) This two-variable distribution also appears to be quasi-normal along both conservation and enrichment axes. The background conservation (*gray* line) for 11-bp is nearly equal to that in Fig. **2-13(a)**; note that the horizontal scale of average phastCons scores is zoomed in by a factor of 2. Averaging conservation over all permutations in this case yields $\bar{s} \pm \sigma_s = 0.398 \pm 0.021$—more than *3 times narrower* than permutations of the *Mef2* motif. Enrichment z-scores have $\bar{z} \pm \sigma_z = 16.11 \pm 2.87$ (with fold-enrichment averaging 2.11 ± 0.20)—rather higher average enrichment than for *Mef2* permutations, but with less variability. Pearson correlation between all enrichment scores and conservation scores for the whole set $\{(s_m, z_m),\ m = 1, \ldots, 277{,}200\}$ yields $r^2 = 0.0410 \pm 0.0008$, i.e., still small but significant. The tilt of the major axis of the cloud just above the horizontal can be discerned in Fig. **2-13(b)**. The unpermuted reference motif BTWTWTHWDDH (circled point) lies among the most highly enriched permutations, but its instances do not have an especially high average phastCons score.

The similarity scores of all the permutations with BTWTWTHWDDH seem to be centered and concentrated fairly symmetrically, except for a slight skew towards the reference permutation (circle). Overall correlation of similarity scores with conservation s-scores in this case is $r_s = 0.1717 \pm 0.0019$ (Pearson-r) for the whole set; correlation of similarity scores with enrichment z-scores is $r_z = 0.2315 \pm 0.0018$. Correlation of similarity to BTWTWTHWDDH with conservation is therefore less for the permutations of *Mef2-HiSco*

than it was among the permutations of YTAWWWWTAR. This is in line with the fact that the *Mef2-HiSco* motif was not chosen by MEME based on conservation at all. It was chosen for its enrichment, and indeed it has a superior enrichment z-score among the set of its permutations, as illustrated by Fig. **2-13(b)**. The correlation of enrichment with degree of similarity to BTWTWTHWDDH is just a little higher than for *Mef2*'s permutations -- probably because *Mef2-HiSco* was optimized by MEME in the first place from a set of related sequences that are present in comparably high numbers in the Mef2 peaks.

Taken together, these results imply that such an *in silico* study of known motifs can shed light on the degree to which they are likely to be functional or merely common. The unusually high conservation rate of instances of the unpermuted *Mef2* motif occurs in parallel with their extremal enrichment, confirming that the specificity of the YTAWWWWTAR sequence plays a part in both its function and frequency. The *Mef2-HiSco* motif, on the other hand, though also highly enriched (its *raison d'être*) may merely be common; its permutations' conservation scores are enhanced above background, but it's not clear that BTWTWTHWDDH in particular stands out as a special enriched *and* conserved member of this set.

(*N.B.* that as a control I also looked at similarity scores measured against randomly chosen permutations of YTAWWWWTAR and BTWTWTHWDDH, but these results are not discussed here.)

Chapter 3

Conclusions

Development and gene regulation in the brain is particularly complex, as most genes are expressed there, the number of cell types is large, and the morphology is intricate. In tackling the molecular and genetic basis of disorders of nervous system development and function, a critical component in the reverse engineering of the mechanisms regulating gene expression is the identification of the set of functional targets of regulatory factors. It is essential to characterize the regulatory network as completely as possible, including the frequency, location, and viability of transcription factor target sites in the genome. In this thesis I have investigated the enrichment and functionality, via comparative genomics, of potential and empirical binding sites of MEF2, a transcription factor known to play a regulatory role in activity-dependent synapse modeling and neuronal function [15, 16].

In conjunction with recent experiments Ref. [1] assessing MEF2-regulated target genes in stimulated rat neurons, I have quantified the relative enrichment of MEF2 binding motifs and their degree of conservation in three regions: the rat genome at large, candidate MEF2 target gene regions probed on a tiling microarray, and specific regions identified as peaks of MEF2 binding. I found the canonical MEF2 binding consensus sequence YTAWWWWTAR (*Mef2*) to be over 10-fold enriched in Mef2 binding peaks. Alternative motifs suggested by MEME purely on account of their enrichment in the Mef2 peaks were about 4-fold enriched in these regions, though only at a level comparable to or slighly higher than the level of the canonical motif *Mef2* (at $\sim 1.3\times$) in the gene regions and genome-wide. All these motifs were present at significantly higher rates than background in every region. The only control motif not merely comparable to background was a random permutation of the *Mef2* sequence.

Conservation of motif instances across the whole rat genome, in the tiled-array gene regions, and in the Mef2 peaks was evaluated using phastCons scores s [26]. High-scoring instances were assumed to have a high degree of functionality with respect to MEF2 binding in rat. I found phastCons scores to be strongly bimodal towards $s \to 0$ and $s \to 1$, in about

a 10:1 ratio in all rat chromosomes. Average background conservation was significanty higher in Mef2 peaks with $\bar{s}_{\text{peaks}} \approx 0.25$. Conservation in the Mef2 peaks themselves was concentrated in islands of high and low conservation; consequently, individual motif instances tended to have either high or low conservation. I confirmed that conservation on average was enhanced near the center of Mef2 peaks. I illustrated these properties concisely as color-coded conservation maps over the peak regions.

Individual instances of the canonical *Mef2* motif were discovered to have an unusually high average conservation score, $\bar{s}_{Mef2} \approx 0.62$, in the Mef2 peaks. I furthermore found that conservation levels that were significantly high above background extended into the neighboring sequence in both directions as far as 100 bp from the motif site—see Fig. **2-10**. Moreover, *Mef2* instances stood out as significantly conserved above backgrounds rates throughout the rat genome, at $\bar{s} \approx 0.25$ *vs.* $\bar{s}_{\text{genes}} \approx 0.10\text{-}0.15$ on average. The *Mef2*-variant motifs' instances also showed some elevation of conservation at their loci and in at least some of the surrounding sequence in Mef2 peaks; the variants *Mef2-HiSco* and *Mef2-var3* and the permutation *Mef2-Perm* captured some of the canonical motif's "extra" conservation outside the peak regions as well, but not nearly as much. A survey of all permutations of *Mef2* and *Mef2-HiSco* established that the particular sequence YTAWWWWTAR was endowed with both unusually high frequency and conservation in Mef2 peaks, whereas the MEME-identified BTWTWTHWDDH was confirmed as enriched but not remarkably conserved. I additionally found a surprising enhancement of conservation near instances of the motif TGACGTMD in the Mef2 peaks, suggesting that the transcription factor CREB might interact with MEF2, either directly or indirectly, in the context of MEF2 binding.

The above characterizations of putative binding sites of *Mef2* and related motifs demonstrate that a straightforward accounting of the frequency and evolutionary persistence of a transcription factor's target genes and target binding sites can guide the interpretation of experimental binding data. High enrichment *per se* of consensus sites is not necessary for a high likelihood of functional binding, but knowledge of enhancement over rates expected from random background is a useful adjunct to DNA binding data and genomic sequence. Although poorly conserved though functional consensus sites (which would be false negatives here) might indicate recently acquired evolutionary function, cataloging highly conserved occurrences as I've done in this work is efficient and can serve to pinpoint specific sites of actual binding and function of important regulatory factors such as MEF2. Biomedical applications of these informatics-based studies will depend on such detailed knowledge of the underpinnings of MEF2 function and that of other regulatory pathways.

Chapter 4

Further Work

The work discussed here on binding motif properties in the rat genome in relation to the neuronal function of the transcription factor (TF) MEF2 can be elaborated and extended in several ways. Here are some gaps and questions that remain to be addressed, and follow-up questions that might be pursued as continuations of the current work, some with more sophisticated methods:

- Sort out the degree of conservation in the control *vs.* non-control genes on the tiled microarray.
- Are the MEF2-controlled genes that do contain Mef2 peaks, and the peaks that do contain *Mef2* consensus sites, specifically enriched for certain Gene Ontology categories pertinent to neurological development and function?
- Can the "neighborhood" effect of conservation near motif instances be modeled by a range of segment lengths having high conservation ($s = 1$ *vs.* $s = 0$)? E.g., 10-bp motifs that fall randomly within many stretches of the same length would accumulate an average conservation score having a purely triangular profile seen near the center of Figs. **2-10**–**2-12**, while a range of high-conservation island lengths could yield even more pinched peaks. Can the distribution of these segments' lengths be used to further characterize the expected affinity of factor binding?
- Investigate in greater detail the phastCons distribution over the *Mef2* motif's whole set of instances rather than just the average of all the instances and their neighborhoods.
- Employ a more sophisticated algorithm for identifying actual peaks of Mef2 binding. An attractive candidate is the Joint Binding Deconvolution (JBD) algorithm developed at M.I.T. [30], which uses ChIP-chip or ChIP-seq data to infer binding factor loci at high resolution.

- Analyze the the location and density of *Mef2* (and other) motif instances with respect to gene boundaries (putative promoter regions, enhancer regions, and 3′UTRs). Is there an association with distances upstream and downstream of transcription start sites (TSS)? With high *vs.* low conservation scores? Is there a higher correlation between the presence of multiple binding sites in a gene in Mef2 binding peaks?

- Analyze the distribution of loci of *Mef2* instances with low conservation score, in Mef2 peaks and in genes, with an eye towards which of these may nevertheless be functional (i.e., false negatives) due to very recent evolutionary divergence.

- Identify specifically highly conserved *Mef2* sites in the genome at large, and their proximity to transcription start sites—distance upstream, presence in promototers, in introns, etc. Are the instances far from TSSs actual sites of Mef2 binding, whether or not MEF2 is activated? Are they highly conserved owing to other functions of MEF2 in other cell types, e.g., muscle progenitor differentiation?

- Extend these analyses of enrichment and conservation to a wide set of candidate motifs (e.g., the whole Xie set [4, 31]), particularly average base-by-base phastCons scores for all instances of each motif as in Fig. **2-12**.

- Analyze ChIP experiments similar to those described here and in Ref. [1] using direct sequencing of chromo-immunoprecipitated fragments (ChIP-seq), and tailor enrichment and conservation analyses to the large amount of data from high-throughput sequencing of ChIP samples.

Most ambitious would be an effort to combine motif-instance and -enrichment data with both TF and TF-target expression data to infer explicit programs of gene expression reflecting causal relations between TFs and their *cis*-acting effects on specific targets. One possible model for this is the thermodynamic model of transcriptional control detailed in Ref. [32]. Another versatile approach has been developed in Ref. [28], which is based on a "phylogenetic framework" of closely-related species. However, the incorporation of both enrichment and *deep* conservation into a motif analysis in a context of factor binding and target expression has yet to be fully realized.

Appendix A

Mathematical Details

In this Appendix I collect miscellaneous mathematical notes pertinent to different topics discussed in the main text. The first section reviews some pertinent features of common statistical distribution used elsewhere in this thesis. General statistical principles are discussed in Ref. [33]. The second section addresses some issues regarding motif frequency counting.

A.1 Statistical Distributions

Most statistical distributions arise from counting—randomly sampling events in some kind of population and approximating formulas for these counts. This generally involves factorials, $n! = \Gamma(n+1)$ for integral $n \geq 0$, which are well approximated by Stirling's formula,

$$n! = \sqrt{2\pi\, n}\; e^{-n}\, n^n \left\{1 + \tfrac{1}{12n} + \mathcal{O}\left(n^{-2}\right)\right\} \tag{A.1}$$

where the "order of" $\mathcal{O}$-notation indicates the upper limit of all remainder terms. Without the $\frac{1}{12n}$ correction Eq. (A.1) is accurate to about 2% for $n \geq 4$ and 4% for $n \geq 8$; with the correction it is already accurate to `0.001` for $n = 1$. The following results will also require this Taylor expansion of the natural logarithm function:

$$\ln(1+\delta) = \delta - \tfrac{1}{2}\delta^2 + \mathcal{O}\left(\delta^3\right), \qquad |\delta| < 1 \tag{A.2}$$

Binomial distributions describe random samples based on proportions with two categorical measurement outcomes of a variable x, one (e.g., Heads) with probability p and the other (Tails) with probability $(1-p)$. For a sample of size N, the number of ways of obtaining Heads H times out of N, and therefore Tails for the remaining T counts, equals the binomial coefficient

$$\binom{N}{H} \equiv \frac{N!}{H!\,T!} \approx \frac{1}{\sqrt{2\pi\, HT/N}}\, \frac{N^N}{H^H\, T^T} \times \left\{1 + \mathcal{O}\left(\frac{1}{N}\right)\right\}, \qquad H + T = N \tag{A.3}$$

using Eq. (A.1) for the factorials. The *probability* of obtaining H Heads and T Tails equals

$$P_N(p;H) = \binom{N}{H} p^H (1-p)^T, \quad H + T = N \qquad \Rightarrow \qquad \sum_{H=0}^{N} P_N(p;H) = 1 \tag{A.4}$$

which, as indicated, is normalized over all possible outcomes for the variable of interest (Heads here) for any p ($0 \leq p \leq 1$). The mean and variance of this distribution are easily shown to equal $\mu = \langle H \rangle = pN$ and $\sigma^2 = \left\langle (H - pN)^2 \right\rangle = Np(1-p)$, respectively, where angular brackets indicate averages. The fact that the standard deviation scales with $\sqrt{N} < N$ establishes the localization or "central tendency" of this distribution, a critical feature of many statistical distributions.

The use of Eqs. (A.1) and (A.2) to approximate the binomial distribution is fundamental to the reduction of all such distributions to essentially gaussian form and their interpretation as normal distributions. It is straighforward to show that, for fixed N and p, Eq. (A.4) is maximized when H equals its mean value pN. Hence, we shift H to a new variable Δ that measure deviations from the mean (from $\Delta = 0$) by writing $H = pN + \Delta$ and $T = (1-p)N - \Delta$ to maintain $H + T = N$. We note that $N = e^{\ln N}$, $\ln pN = \ln p + \ln N$, etc.,

and assume that extra terms $\sim \mathcal{O}(\Delta/N)$ are negligible compared to unity. Then combining Eqs. (A.2)–(A.4) yields

$$P_N(p;H) = \frac{1}{\sqrt{2\pi\, Np(1-p)}}\, \exp\Big[-\Big(\cdots\Big)\Big] \times \left\{1 + \mathcal{O}\left(\frac{\Delta}{N}\right)\right\} \tag{A.5}$$

where

$$\begin{aligned}\Big(\cdots\Big) &= -N\ln N + (pN+\Delta)\Big[\ln(pN+\Delta) - \ln p\Big] \\ &\qquad + ([1-p]N-\Delta)\Big[\ln([1-p]N-\Delta) - \ln[1-p]\Big] \\ &= (pN+\Delta)\ln\left(1+\frac{\Delta}{pN}\right) + ([1-p]N-\Delta)\ln\left(1-\frac{\Delta}{[1-p]N}\right) \\ &= \frac{1}{2}\frac{\Delta^2}{pN} + \frac{1}{2}\frac{\Delta^2}{(1-p)N} + \mathcal{O}\left(\frac{\Delta^3}{N^2}\right) = \frac{1}{2}\frac{\Delta^2}{p(1-p)N} + \mathcal{O}\left(\frac{\Delta^3}{N^2}\right)\end{aligned} \tag{A.6}$$

Taken together Eqs. (A.5) and (A.6) produce a ***gaussian*** in the variable $\Delta = H - pN$:

$$P_N(p;H) = \frac{1}{\sqrt{2\pi\,\sigma^2}}\, \exp\left[-\frac{(H-\mu)^2}{2\,\sigma^2}\right] \times \left\{1 + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right\}, \quad \begin{cases} \mu = pN \\ \sigma^2 = Np(1-p) \end{cases} \tag{A.7}$$

Since this function is exponentially small unless $\Delta^2 \sim \mathcal{O}(N)$, the error in this expression is shown as only $\sim \mathcal{O}\left(N^{1/2}/N\right)$ at most and is likely even smaller. Note that this approximation to the binomial distribution is still normalized and has unchanged variance $\sigma^2 \propto N$.

The shifted and rescaled variable $(H-\mu)/\sigma$ is often called 'z', which becomes the argument of a standard gaussian, $(2\pi)^{-1/2}e^{-z^2/2}$, with zero mean and unit standard deviation. As usual, the one-tailed p-value for a given score z under this distribution is defined as the cumulative probability of observing this *or any more extreme* score: $p(z) = \frac{1}{\sqrt{2\pi}}\int_z^\infty dz\, e^{-z^2/2}$.

A ***Poisson distribution*** is a special case of the binomial distribution for rare events ($p \approx 0$) that have many opportunities to occur ($N \gg 1$) in such a way that the average number of random occurrences $\mu = pN \geq 0$ is finite. Without reproducing the derivation here, we note the simple result:

$$P(\mu;n) = e^{-\mu}\,\frac{\mu^n}{n!}, \qquad \begin{cases} \mu = pN \\ \sigma^2 = pN = \mu \end{cases} \tag{A.8}$$

where $n = 0,\ 1,\ 2,\ \ldots$ is the number of events observed *vs.* the expected number μ. The mean of this distribution indeed equals $\langle n\rangle = \mu$ and its variances *also* equals $\sigma^2 = \mu$. For not-too-small μ, the Poisson distribution is still well approximated by a gaussian; moreover, fluctuations in observed values of n, of order $\sigma \approx \sqrt{\mu}$, decrease with respect to the mean observed value as $\Delta n/\langle n\rangle \sim \sigma/\mu = 1/\sqrt{\mu} = 1/\sqrt{\langle n\rangle}$.

For binomial (and multinomial) distributions, it is assumed that the probability of each kind of event is fixed. This is equivalent to assuming that there is an essentially infinite pool of mixed outcomes, e.g., Heads and Tails events, present in the pool in the ratio $p/(1-p)$. It doesn't matter whether samples of size N are drawn with replacement or not because the pool is so large: Heads will continue to come up with probability p. If, however, the pool has a finite size N_{total}, it may be of interest to calculate the ***hypergeometric distribution***—the exact probability of finding a certain number of Heads in a sample of size N drawn *without replacement*, given the specific numbers of Heads and Tails in the whole pool. For example, continuing with Heads/Tails categories, suppose the pool contains N_{Heads} and N_{Tails} of these types of events, with total $N_{\text{total}} = N_{\text{Heads}} + N_{\text{Tails}}$. This specifies probabilities $p = N_{\text{Heads}}/N_{\text{total}}$ for randomly getting Heads once or $1-p = N_{\text{Tails}}/N_{\text{total}}$ for getting Tails once. In a random sample of size N, some number H out of N actually turn up Heads—which must come from the sub-pool of N_{Heads} Heads. The remainder of the sample, $T = N - H$ out of N, must turn up Tails—and these must come from the remaining sub-pool of $N_{\text{Tails}} = N_{\text{total}} - N_{\text{Heads}}$ Tails. Considering all the ways N could have been selected from the entire pool of N_{total}, leaving $N' = N_{\text{total}} - N$ *un*selected, the probability of obtaining H Heads equals this ratio of binomial coefficients:

$$P(N_{\text{total}}; N_{\text{Heads}}, N; H) = \frac{\binom{N_{\text{Heads}}}{H} \times \binom{N_{\text{Tails}}}{T}}{\binom{N_{\text{total}}}{N}} \tag{A.9}$$

$$= \frac{N_{\text{Heads}}!\, N_{\text{Tails}}!\, N!\, N'!}{N_{\text{total}}!\, H!\, (N_{\text{Heads}} - H)!\, T!\, (N_{\text{Tails}} - T)!}$$

In a random sample of size N, it is easy to show that on average the fraction of the sample that turns up Heads still equals p; i.e., the mean number of Heads equals $\mu = \langle H \rangle = pN = N_{\text{Heads}}N/N_{\text{total}}$. With a little more work, one can also show that the variance of H equals

$$\sigma^2 = \frac{N_{\text{Heads}} N_{\text{Tails}} N N'}{N_{\text{total}}^2 (N_{\text{total}} - 1)} \approx N_{\text{total}}\, p(1-p) f(1-f)\,, \qquad p \equiv \frac{N_{\text{Heads}}}{N_{\text{total}}}\,, \quad f \equiv \frac{N}{N_{\text{total}}} \tag{A.10}$$

which scales linearly with N as before.

In fact, the hypergeometric distribution, Eq. (A.9), can also be approximated by a gaussian in the variable H as before, with precisely this mean and variance. To calculate an exact p-value (not to be confused with the probability p here!) for drawing H Heads in a sample of size N from a finite pool of Heads and Tails, one could just add up the hypergeometric probabilities $P(N_{\text{total}}; N_{\text{Heads}}, N; H')$ for values $H' = H$ through N (or N_{Heads}, if $N_{\text{Heads}} < N$). This is called ***Fisher's exact test***. The gaussian approximation to Eq. (A.9) may also be employed, but then a Yates correction $(H - \mu) \to |H - \mu| - \frac{1}{2}$ is required for

small N for turning the sum into an integral. Generalizations from bivalent to multivalent variables, and from binomial to multinomial distributions, are straightforward; however, more complicated versions of the hypergeometric probabilities are best approximated as a χ^2 test as follows.

The gaussian form of Eq. (A.9) stems from the behavior of the four H- and T-dependent factorials, each of which itself has a characteristic gaussian dependence. Once again write $H = \langle H \rangle + \Delta$ in terms of a deviation Δ from the expected value $\langle H \rangle = pN = pfN_{\text{total}}$, where $f = N/N_{\text{total}}$ is the fraction of the total sampled. Similarly, the other factors are written as shifts from their expected values: $T = [1-p]fN_{\text{total}} - \Delta$, $N_{\text{Heads}} - H = p[1-f]N_{\text{total}} - \Delta$, $N_{\text{Tails}} - T = [1-p][1-f]N_{\text{total}} + \Delta$, where each term $\pm\Delta$ is required by the given subset totals. Each of the four gaussians has an exponent of the form $(H - \langle H \rangle)^2 / 2\langle H \rangle$, etc. Then the product of four gaussians has four exponents that can be summed together:

$$
\begin{aligned}
&P(N_{\text{total}}; N_{\text{Heads}}, N; H) \\
&\quad \propto \exp\left[-\frac{1}{2}\frac{(H - pfN_{\text{total}})}{pfN_{\text{total}}} - \frac{1}{2}\frac{(T - [1-p]fN_{\text{total}})}{[1-p]fN_{\text{total}}} \right. \\
&\qquad \left. -\frac{1}{2}\frac{([N_{\text{Heads}} - H] - p[1-f]N_{\text{total}})}{p[1-f]N_{\text{total}}} - \frac{1}{2}\frac{([N_{\text{Tails}} - T] - [1-p][1-f]N_{\text{total}})}{[1-p][1-f]N_{\text{total}}} \right] \\
&\quad = \exp\left[-\frac{1}{2}\frac{\Delta^2}{\sigma^2} \right], \qquad \Delta = H - pfN_{\text{total}}, \quad \sigma^2 = N_{\text{total}}\, p[1-p]f[1-f]
\end{aligned}
\tag{A.11}
$$

The point is that more complicated sampling patterns have the same structure. For example, suppose a data set is partitioned into $r \geq 2$ sampling sets of size N_i that are fractions $f_i/N_i/N_{\text{total}}$ of the whole set ($i = 1,\ 2,\ \ldots,\ r$), and if p_j ($j = 1,\ 2,\ \ldots, c$) is the probability of measuring category $\#j$ out of $c \geq 2$ possible values $\{y_j\}$. The expected value of seeing a member of subset $\#i$ with value $\#j$ is $E_{ij} = f_i p_j N_{\text{total}}$ but in general any value O_{ij} might be observed. The analog of Eqs. (A.9) and (A.11) for the probability of seeing a set of observations $\{O_{ij}\}$ is then a gaussian whose exponent is a sum of $r \times c$ terms of all of the same form $(O - E)^2/E$:

$$
P(\{O_{ij}\}) = \exp\left[-\frac{1}{2}\chi^2 \right], \qquad \chi^2 \equiv \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad E_{ij} = f_i p_j N_{\text{total}}
\tag{A.12}
$$

This is of course the original of the ***chi-squared*** statistic (χ^2). Samples of independent, identically and normally distributed random variables have a χ^2 distribution of sum-of-squared-deviations. Other common statistical distributions based on samples of normal, random variables are t-tests and the F-statistic used in analysis of variance (ANOVA).

A.2 Motif Frequencies

The enrichment of any set of sequences in any region of the genome has to measured against some kind of background, or null hypothesis. Unless specified otherwise, I take the background to be *random* occurrences of any and all members of the set.

The frequency of random occurrences of a DNA consensus sequence such as the *Mef2* motif YTAWWWWTAR can be estimated if one assumes an infinitely long string of bases {ACGT} distributed randomly with prescribed average probabilities $\{p_A, p_C, p_G, p_T\}$. For given CG content p_{CG} and $p_A+p_C+p_G+p_T = 1$, these proportions are fixed by $p_C = p_G = \frac{1}{2}p_{CG}$ and $p_A = p_T = \frac{1}{2}p_{AT} = \frac{1}{2}(1 - p_{CG})$. For 2-base positions, since S = {C, G} and W = {A, T} we have $p_S = p_{CG}$ and $p_W = p_{AT} = 1 - S$ but it always holds that $p_R = p_{AG} = p_A + p_G = \frac{1}{2}$ and likewise $p_Y = p_{CT} = \frac{1}{2}$ and $p_M = p_{AC} = \frac{1}{2}$ and $p_K = p_{GT} = \frac{1}{2}$.

Taking *Mef2* under a uniform distrubution ($p_{CG} = \frac{1}{2}$) as an example, the single bases A and T each occur on each strand with probability $\frac{1}{4}$ while the 2-base positions Y, W, and R each occur on each strand with probability $\frac{1}{2}$. The whole motif therefore occurs on each DNA strand with probability

$$p_Y \cdot p_T \cdot p_A \cdot p_W \cdot p_W \cdot p_W \cdot p_W \cdot p_T \cdot p_A \cdot p_R = \tfrac{1}{2} \cdot \tfrac{1}{4} \cdot \tfrac{1}{4} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{4} \cdot \tfrac{1}{4} \cdot \tfrac{1}{2} = \left(\tfrac{1}{4}\right)^4 \times \left(\tfrac{1}{2}\right)^6 = \left(\tfrac{1}{2}\right)^{14}$$

i.e., at 1 in $2^{14} = 16{,}384$ base positions. Owing to the double ambiguities of Y = {C, T}, W = {A, T}, and R = {A, G}, there are $1^4 \times 2^6 = 64$ different realizations of this motif, such as CTATATATAA and GTAAAAATAA—all equally probable. Thus, another way of obtaining the same frequency is to note that these 10-bp sequences arise randomly as a fraction $\frac{64}{4^{10}} = 4^{-7}$ of all possible 10-bp sequences, the same as above.

The *Mef2* motif is *palindromic*—it equals its own reverse complement—so every time one of its realizations occurs on one DNA strand (POS, say) another one coincidentally occurs on the complementary strand (NEG). There is no additional information obtained from the second strand, no extra "search space" for counting motif instances. However, a *non*palindromic motif might be independently searched on both DNA strands without encountering this redundancy. As a simple illustration consider the palindromic toy motif AWT: its 2 possible matches on the forward direction of either strand are AAT and ATT. However, the reverse complement of this set is the same, so this motif is found at only 2 of every $4^3 = 64$ bases (if $p_{CG} = \frac{1}{2}$). As a nonpalindromic example, the YWT has 4 possible matches on the POS strand: CAT, CTT, TAT, TTT. These also occur on the forward direction on the NEG strand, which would appear as their reverse complements on the POS strand: ATG, AAG, ATA, AAA, respectively. Since these 8 sequences are all different, there are in effect twice as many loci at which one would randomly find this motif on double-stranded DNA as there would be on a single strand; its frequency equals $\frac{8}{64} = \frac{1}{8}$, not $\frac{4}{64} = \frac{1}{16}$. On the other hand, the motif WWT has forward matches AAT, ATT, TAT, TTT, and reverse

complement matches ATT, AAT, ATA, AAA. This motif can have 6 different matches: its frequency $\frac{6}{64}$ is more than the $\frac{4}{64}$ that would be found on one strand, less than the $\frac{8}{64}$ that would be found on both strands if there were no redundancy. The only way to precisely calculate motif frequencies for random DNA, therefore, is to collect all the *unique* possible matches of the motif and its reverse complement, calculate the probability of each match, and add these up to obtain the motif's total probability. Only for a perfectly palindromic motif is the reverse-complement accounting unnecessary.

Returning to the *Mef2* motif, the CG content in the rat genome as a whole equals $p_{\mathtt{CG}} = 0.418$ (relatively AT-rich) while in the Mef2 peaks discussed in Sec. **2.2** it equals $p_{\mathtt{CG}} = 0.489$ (close to uniform). In the latter regions, $p_{\mathtt{W}} = 0.5110$, $p_{\mathtt{A}} = p_{\mathtt{T}} = 0.2555$, and $p_{\mathtt{Y}} = p_{\mathtt{R}} = \frac{1}{2}$. Thus, the total probability for all 64 matches to YTAWWWWTAR equals

$$p_{\text{Mef2}} = (0.2555)^4 \cdot (0.5110)^4 \cdot (0.5000)^2 = 7.264 \times 10^{-5} \approx \frac{1}{13{,}766} \qquad \text{(Mef2 peaks)}$$

which implies that this motif would occur slightly more often *by chance* than it would for 50% CG content. In the AT-richer whole rat genome, on the other hand, with $p_{\mathtt{W}} = 0.582$, the probability of finding the canonical *Mef2* motif goes up significantly:

$$p_{\text{Mef2}} = (0.2910)^4 \cdot (0.5820)^4 \cdot (0.5000)^2 = 2.057 \times 10^{-4} \approx \frac{1}{4{,}862} \qquad \text{(rat genome)}$$

The occurrence of a match to the *Mef2* motif is a rare, random event and is thus amenable to Poisson statistics. If N base pairs are available, then the mean number of matches one can expect to observe per base with probability p in a random sequence of bases simply equals $N_{\text{expect}} = p \times N$. The variance equals the mean in a Poisson process so the expected numbers under the null hypothesis are given in Table **2.1** as $pN \pm \sqrt{pN}$, i.e., as mean expected value with an "error" of one standard deviation.

The Mef2-variant motifs found by MEME (Sec. **2.2**) were nonpalindromic. Their background frequency and expected number in each region of interest were calculated in the same as way as described here and are also shown in Table **2.1**. Enrichment scores shown depend on the number N_{found} of instances of each motif found in each region. The enrichment itself is simply the ratio $N_{\text{found}}/N_{\text{expect}}$. The enrichment z-score equals the difference of the measured number of instances with respect to the expected number, in units of Poissonian standard deviations: $z = (N_{\text{found}} - N_{\text{expect}}) / \sqrt{N_{\text{expect}}}$.

A.3 Similarity Scores

In Sec. **2.4.3** permutations of the motifs *Mef2* and *Mef2-HiSco* were each compared to their unpermuted, "reference" sequences YTAWWWWTAR and BTWTWTHWDDH with regard to their enrichment and conservation properties in Mef2 peaks. The question arose of how these properties might vary from that of the reference motif when the permuted sequence was more or less similar to the unpermuted sequence. The method I present here for similarity score between motifs is adapted from an algorithm described in the online Supplementary Information material (section on "Motif clustering") accompanying Ref. [4].

Two sequences of nucleotides S_1 (length n_1) and S_2 (length n_2) are to be compared for similarity of their base content and for the alignment their sequence that optimizes that comparison, given a background base content. The background β is simply determined by its CG content; e.g., $p_{\mathsf{CG}} = 0.4892$ for the Mef peaks implies the ordered list of 4 probabilities

$$\beta = (p_{\mathsf{A}}, p_{\mathsf{C}}, p_{\mathsf{G}}, p_{\mathsf{T}}) = (0.255, 0.245, 0.245, 0.255) \tag{A.13}$$

A *flat* background has equally weighted bases:

$$\beta_0 = (0.25, 0.25, 0.25, 0.25) \tag{A.14}$$

Bases at individual positions are specified by the usual alphabet

$$\begin{aligned} &\mathsf{A},\ \mathsf{C},\ \mathsf{G},\ \mathsf{T}, \\ &\mathsf{M} = \{\mathsf{A},\mathsf{C}\},\ \mathsf{K} = \{\mathsf{G},\mathsf{T}\},\ \mathsf{R} = \{\mathsf{A},\mathsf{G}\},\ \mathsf{Y} = \{\mathsf{C},\mathsf{T}\},\ \mathsf{W} = \{\mathsf{A},\mathsf{T}\},\ \mathsf{S} = \{\mathsf{C},\mathsf{G}\}, \\ &\mathsf{B} = \{\mathsf{C},\mathsf{G},\mathsf{T}\},\ \mathsf{D} = \{\mathsf{A},\mathsf{G},\mathsf{T}\},\ \mathsf{H} = \{\mathsf{A},\mathsf{C},\mathsf{T}\},\ \mathsf{V} = \{\mathsf{A},\mathsf{C},\mathsf{G}\}, \\ &\mathsf{N} = \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\} \end{aligned} \tag{A.15}$$

which allows for zero, double, triple, or complete ambiguity. Each motif base can also be represented as a list of probabilities, e.g., $(0,0,1,0)$ for 'G', $(\frac{1}{2},0,\frac{1}{2},0)$ for 'R', $(\frac{1}{3},\frac{1}{3},0,\frac{1}{3})$ for 'H', and $\beta_0 = (\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})$ for 'N'. Each set is given equally weight probabilities in a motif since I treated ambiguated letters equally when searching for motif instances in genomic sequence. An entire sequence S can thus be represented as a *position-weight matrix* $\mathbf{M}$. If S needs to be extended at either end by nonspecific sequence, a string of background N's should be used with weights β, Eq. (A.13), appropriate to the region in which motif instances are to be found. For example, the motifs *Mef2* and *Mef2-HiSco* embedded in background sequence look like

$\mathbf{M}_{Mef2} =$

$\cdots\ \beta$	β	Y	T	A	W	W	W	W	T	A	R	β	$\beta\ \cdots$
$\cdots$	0.255	0	0	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	1	$\frac{1}{2}$	0.255	$\cdots$
$\cdots$	0.245	$\frac{1}{2}$	0	0	0	0	0	0	0	0	0	0.245	$\cdots$
$\cdots$	0.245	0	0	0	0	0	0	0	0	0	$\frac{1}{2}$	0.245	$\cdots$
$\cdots$	0.255	$\frac{1}{2}$	1	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	0	0	0.255	$\cdots$

(A.16)

and

$\mathbf{M}_{Mef2\text{-}HiSco} =$

$\cdots\ \beta$	β	B	T	W	T	W	T	H	W	D	D	H	β	$\beta\ \cdots$
$\cdots$	0.255	0	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.255	$\cdots$
$\cdots$	0.245	1	0	0	0	0	0	$\frac{1}{3}$	0	0	0	$\frac{1}{3}$	0.245	$\cdots$
$\cdots$	0.245	1	0	0	0	0	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0.245	$\cdots$
$\cdots$	0.255	1	1	$\frac{1}{2}$	1	$\frac{1}{2}$	1	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.255	$\cdots$

(A.17)

To score two sequences S_1 and S_2 for similarity, they must be lined up and compared base-for-base. If two bases have the same composition, e.g., both G's or both R's, then their contribution to the score should be maximal; orthogonal choices such as G *vs.* W should contribute zero. Considering all aligned pairs of bases, this implies that the *correlation* over all elements of their weight matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ is a suitable measure of similarity. *The similarity score λ is defined as the value of r for the alignment shift a that maximizes the Pearson correlation r between $\mathbf{M}_1$ and $\mathbf{M}_2$ in a background β:*

$$\lambda\,(S_1, S_2; \beta; a) \equiv \arg\max_a \frac{\sum_i \sum_b \mathbf{M}_1(i,b)\,\mathbf{M}_2(i-a,b) - \mu_1\,\mu_2}{\sigma_1\,\sigma_2}\,, \tag{A.18}$$

where μ_k is the average of all (non-background) elements of $\mathbf{M}_k$ and σ_k their standard deviation. Here $\mathbf{M}_k(b,i)$ denotes the weight of base $b \in \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$ (i.e., row of $\mathbf{M}_k$) at position i (column) in sequence $k = 1$ or 2. The shift a is the number of bases that S_2 is offset from alignment with S_1 at their leftmost bases. The sum $\sum_i$ extends over all base positions of S_1 and S_2 whether they overlap with one another or with the flanking background.

By way of example, Table **A.1** shows the optimized alignment and similarity score λ

(comparison to YTAWWWWTAR) for each of the *Mef2*-variant and control motifs discussed in Chapter **2**:

Mef2 vs. ...	Shift a	λ	o o o o Y T A W W W W T A R o o
Mef2	0	1.000	o o o o Y T A W W W W T A R o o
Mef2-var1	0	0.803	o o o o H W A W A W W W A R o o
Mef2-var2	-2	0.524	o o K B Y T D T T T W D D o o o
Mef2-var3	-1	0.732	o o o D R T W W T T W T A R o o
Mef2-HiSco	0	0.595	o o o o B T W T W T H W D D H o
Mef2-Perm	-1	0.668	o o o W Y W A A T R W T W o o o
Mef2-Rdm1	2	0.454	o o o o o o o A K C T W W A G M T
CREB	-4	0.183	T G A C G T M D o o o o o o o o o
BACKGROUND	.....	0.019	o o o o o o o o o o o o o o o o o

Table A.1. Example similarity scores for the *Mef2* motif *vs.* other motifs listed in Table **2.1**. Optimal shift a and resulting score λ. Motifs shifted with respect to *Mef2* for optimal score. Here 'o' represents the background weights β for Mef2 peaks.

Appendix B

Computer Codes

In this Appendix I summarize the major computational chores I tackled to carry out the quantitative work described in Chapters **1**–**4**. Scripts and their output are not included here but I note their essential features. I wrote most of the computer code in either `perl v5.8.8` (mostly for sequence mining) or `R v2.4.1` (mostly for statistical analyses and image production).

Most of the bioinformatic analyses that I performed for this thesis required as input either the reference assemblies of entire genomes, alignments among several genomes, or phastCons (conservation) scores based on an alignment. Genomic sequences included both raw sequence from whole chromosome reference contigs as well as annotated genes and gene features from gene reference sequences (RefSeqs). I did not perform any alignments (cf. Ref. [6]) or phylogeny-based conservation analyses (cf. Ref. [25]) myself, but did directly analyze genomic DNA sequence.

The rat genome used in Chapter **2** was *rn4*, the most recent assembly (Nov. 2004) available from UCSC and NCBI. Genomic loci of annotated gene features—whole genes, reference mRNA transcripts, untranslated regions (UTRs), coding exons (CDSs), and pseudogenes—were provided in a single compressed file, "`seq_gene.md.gz`" ($\sim$ 10 MB), downloadable from Ref. [18]. To process this file, I wrote a `perl` script, "`GeneFeatures EXs+UTRs from RefSeqs_2.pl`", to organize its reference assembly entries hierarchically by gene, RefSeq, and gene feature (`5pUTR`, `3pUTR`, `EXon`, and `INtron`, principally). Output contained annotations such as GeneIDs, RefSeq numbers, within-gene feature numbers (1$^{\text{st}}$ exon, 2$^{\text{nd}}$ exon, 5$^{\text{th}}$ intron, ...), start and stop loci, etc., for the whole genome. All gene loci for input to subsequent analyses were drawn from this database. If necessary, any number of `(Start, Stop)` ranges for multiple sets of genes, RefSeqs, and gene features could be combined with arbitrary logical specifications (union, intersection, exclude, etc.) using another `perl` script, "`Ranges from Features_2.pl`".

I downloaded the nucleotide sequence of the whole rate genome from Ref. [22]. In order

to collect arbitrary DNA subsequences for tiled-array genes and Mef2 peaks (Sec. **2.1**), I wrote a `perl` script "`Get_Genomic_Sequences.pl`" to retrieve the sequence for any number of given `(Start, Stop)` loci pairs. An option to include an arbitrary number of "padding" nucleotides on either end of each requested range was useful for extending the annotated loci of genes used in the microarray experiments [1] by 10 kb as input to further bioinformatics. These genomic subregions could then be easily swept for, e.g., CG content.

Exhaustive lists of matches to individual short sequence patterns—such as the *Mef2* motif YTAWWWWTAR and other motifs—needed to be extracted both from genomic sequence (*rn4* chromosomes) and from within sequences delineated by arbitrary loci. The latter included sets of `(Start, Stop)` ranges representing, alternatively, every tiled-array gene or every Mef2 peak (Sec. **2.2**). To find all genomic instances of single motifs on all or any subset of rat chromosomes, I wrote a `perl` script "`Find_Genomic_Motif_Instances.pl`". E.g., for YTAWWWWTAR, blocks of genomic sequence of length 20,000 bp were searched via the 10-bp pattern /[CT][T][A][AT][AT][AT][AT][T][A][AG]/. (A nonzero number of matches to genomic bases marked 'N' was an option but generally disallowed by specifying zero.) An option for outputting padding on either side with every 10-bp genomic sequence match could be used to characterize the sequence surrounding each "hit" if necessary. Another `perl` script, "`Find_Motifs_in_Sequences.pl`", found all instances of single motifs in the set tiled-array gene sequences (including the extra ±10 kb) and in the set of Mef2 peak sequences. Both of these programs also reported the sequence of each matching instance, from which more specific motif position-weight matrices and "logos" [19] could be derived if desired.

For the conservation analyses of Secs. **2.3** and **2.4**, I downloaded phastCons scores for *rn4*, based on an alignment of 8 other vertebrate genomes to the rat genome, from Ref. [26]. These scores were needed for analyzing the conservation of motif instances over whole rat chromosomes (or chromosomal segments), in the tiled-array gene regions, and in the Mef2 peaks, as well as for the average scores of bases in each region. I wrote a `perl` script "`PhastCons_Scores_for_Ranges_3.pl`" to extract scores covering lists of `(Start, Stop)` loci of interest—i.e., for all motif instances already identified in the genome using "`Find_Genomic_Motif_Instances.pl`" or for all instances in gene regions or in Mef2 peaks identified using "`Find_Motifs_in_Sequences.pl`". Output included not only the conservation scores but also the number of input sequence bases that actually had phastCons scores and each range's mean score. Once again, a padding option allowed me to collect scores for 200-bp "neighborhood" sequences around each instance, averages of which are displayed in Figs. **2-9** *et seq.*

Note that all of the genome-mining programs mentioned here are designed to work with any downloaded genome, not just rat. In addition, for work involving the genomic loci of many different motifs, my colleague Dr. J.M. Gray and I have each written code to efficiently

produce this much larger data set all at once rather than one motif at a time. This was required for our studies of the whole Xie set of motifs [4] mentioned in Chapter **1**.

All analyses of phastCons scores, including calculations of average scores in chromosomes, tiled-array gene regions, and Mef2 peaks, were performed using scripts I wrote in `R`. All the images in Chapter **2**, except Fig. **2-1**, were also created using `R`. In particular, the conservation maps of Sec. **2.3.2** involved assembling all conservation scores for all bases in the Mef2 peaks, transforming them in various ways with respect to their sequences' alignment (e.g., centered *vs.* stretched), and color-coding their values. I wrote `R` scripts to collect and average scores over all motifs' instances and their 400-bp neighborhoods in each type of region, and to produce the corresponding plots, in Secs. **2.4.1** and **2.4.2**. The many permuted motifs based on YTAWWWWTAR or BTWTWTHWDDH in Sec. **2.4.3** were enumerated using `R`. I scored each permutation for similarity to *Mef2* or *Mef2-HiSco* in `R`, then scored each for enrichment and conservation in the Mef2 peaks by calling the above mentioned `perl` scripts from `R` and assembling z, s, and λ values into Fig. **2-13(a)**–**(b)** in `R`. I also used `R` for the statistical analyses of Table **2.1**, Sec. **2.4.2**, and elsewhere.

I used Paint Shop Pro `v7.04` to combine the UCSC Genome Browser track images with separate Mef2 peak track images and the color-coded conservation bars in Fig. **2-1**, as well as to add labels to images used as subfigures.

The largest computing jobs (genomic motif searches, phastCons score collection) were run on Harvard Medical School's orchestra cluster. Otherwise, all computations were performed on either a Dell Dimension 9150 work station or a Toshiba Tecra M7 laptop; no job took more than a few hours to run on 2–3-GHz Intel x86 dual-core processors.

This manuscript was typeset using TEX and LATEX based on the `mitthesis.cls` document class available online through the Institvte.

References

[1] S.W. Flavell, T.-K. Kim, J.M. Gray, D.A. Harmin, E. Markenscoff-Papadimitriou, D.M. Bear, and M.E. Greenberg. Genome-wide analysis of the target genes of MEF2, a transcription factor that coordinates synapse development. Submitted to *Neuron*, March 2008.

[2] E.S. Lein et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445:168–176, 2007.

[3] P. Carninci et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Gen.*, 38:626–635, 2006.

[4] X. Xie, J. Lu, 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 mammals. *Nature*, 434:338–345, 2005.

[5] T.L. Bailey, N. Williams, C. Misleh, and W.W Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucl. Acids Res.*, 34:W369–W373, 2006.

[6] UCSC human hg18 17-way multiple alignment downloads (URL). `http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz17way/`.

[7] W. Miller et al. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.*, 17:1797–1808, 2007.

[8] UCSC human hg18 28-way multiple alignment downloads (URL). `http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz28way/`.

[9] S.R. Eddy. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.*, 3:95–102, 2005.

[10] M. Piqué, J.M. López, S. Foissac, R. Guigó, and R. Méndez. A combinatorial code for CPE-mediated translational control. *Cell*, 132:434–448, 2008.

[11] E.K. White, T. Moore-Jarrett, and H.E. Ruley. PUM2, a novel murine puf protein, and its consensus RNA-binding site. *RNA*, 7:1855–1866, 2001.

[12] Allen Brain Atlas (URL). `http://www.brain-map.org/welcome.do`.

[13] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[14] Random forests home page (URL).
`http://stat-www.berkeley.edu/users/breiman/RandomForests/`.

[15] Z. Mao, A. Bonni, F. Xia, M. Nadal-Vicens, and M.E. Greenberg. Neuronal activity-dependent cell survival mediated by transcription factor MEF2. *Science*, 286:785–790, 1999.

[16] S. W. Flavell, C. W. Cowan, T. K. Kim, P. L. Greer, Y. Lin, S. Paradis, E. C. Griffith, L. S. Hu, C. Chen, and M. E. Greenberg. Activity-dependent regulation of MEF2 transcription factors suppresses excitatory synapse number. *Science*, 311:1008–1012, 2006.

[17] UCSC Genome Browser (URL). `http://genome.ucsc.edu/`.

[18] Rat rn4 reference sequence assembly genetic features (URL).
`ftp://ftp.ncbi.nlm.nih.gov/genomes/R_norvegicus/mapview/`.

[19] G.E. Crooks, G. Hon, J.M. Chandonia, and S.E. Brenner. WebLogo: A sequence logo generator. *Genome Res.*, 14:1188–1190, 2004.

[20] R. Pollock and R. Treisman. Human SRF-related proteins: DNA-binding properties and potential regulatory targets. *Genes Dev.*, 5:2327–2341, 1991.

[21] Y.-T. Yu, R.E. Breitbart, L.B. Smoot, Y. Lee, V. Mahdavi, and B. Nadal-Ginard. Human myocyte-specific enhancer factor 2 comprises a group of tissue-restricted MADS box transcription factors. *Genes Dev.*, 6:1783–1798, 1992.

[22] Rat rn4 reference genome (URL).
`http://hgdownload.cse.ucsc.edu/goldenPath/rn4/chromosomes/`.

[23] T.L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. In C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak, editors, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 21–29, Menlo Park, CA, July 1995. AAAI Press.

[24] TRANSFAC transcription factor database (URL).
`http://www.biobase-international.com/pages/index.php?id=transfac`.

[25] A. Siepel et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15:1034–1050, 2005.

[26] Rat rn4 phastcons scores (URL). `ftp://hgdownload.cse.ucsc.edu/goldenPath/rn4/phastCons9way/`.

[27] M. Blanchette et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, 14:708–715, 2004.

[28] P. Kheradpour, A. Stark, S. Roy, and M. Kellis. Reliable prediction of regulator targets using 12 *drosophila* genomes. *Genome Res.*, 17:1919–1931, 2007.

[29] H.D. Youn, T.A. Chatila, and J.O. Liu. Integration of calcineurin and MEF2 signals by the coactivator p300 during T-cell apoptosis. *Embo. J.*, 19:4323–4331, 2000.

[30] Q. Yuan, A. Rolfe, K.D. MacIsaac, G.K. Gerber, D. Pokholok, J. Zeitlinger, T. Danford, R.D. Dowell, E. Fraenkel, T.S. Jaakkola, R.A. Young, and D.K. Gifford. High-resolution computational models of genome binding events. *Nature Biotech.*, 24:963–970, 2006.

[31] X. Xie, T.S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis, and E.S. Lander. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *PNAS*, 104:7145–7150, 2007.

[32] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, 451:535–541, 2008.

[33] S.A. Glantz. *Primer of Biostatistics.* McGraw-Hill Medical, sixth edition, 2005.