INTEGRATING EPIGENETIC PRIORS FOR IMPROVING COMPUTATIONAL IDENTIFICATION OF TRANSCRIPTION FACTOR BINDING SITES

AFFAN SHOUKAT

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

SEPTEMBER 2015

Abstract

Transcription factors and histone modifications play critical roles in tissue-specific gene expression. Identifying binding sites is key in understanding the regulatory interactions of gene expression. Naive computational approaches uses solely DNA sequence data to construct models known as Position Weight Matrices. However, the various assumptions and the lack of background genomic information leads to a high false positive rate. In an attempt to improve the predictive performance of a PWM, we use a Hidden Markov Model to incorporate chromatin structure, in particular histone modifications. The HMM captures physical interactions between distinct HMs. Indeed, the integration of sequence based PWM models and chromatin modifications improve the predictive ability of the integrative model.

Acknowledgements

To begin I would like to thank my advisor, Dr Jorg Grigull for his unending guidance and support. His encouragement and constant mentoring has made this project successful.

Thank you to my committee members Dr Seyed Moghadas and Dr Jane Heffernan. I am grateful to them for their invaluable time, comments, questions, and suggestions regarding my research.

I will always be thankful to my parents Ali and Naila and to my sister Irsa. They constantly pushed me and exposed me to the best educational opportunities while sacrificing their own comfort.

Table of Contents

List of Tables

List of Figures

1 Introduction

There is a growing need for the application of mathematical sciences to biological processes and datasets. In particular, the interdisciplinary field of bioinformatics is such that it combines frameworks, toolkits, and methodologies from mathematics, computer science, and biology to enable analysis of large biological datasets. Most people involved in scientific research are forced to apply concepts of mathematical modeling in order to understand and elucidate biological processes. Mathematical models are the logical extension to the wet-lab methods enabling exploration of complex systems while reducing cost. Recently, there has been much interest in modeling epigenetic mechanisms. In this thesis, we aim to illustrate and characterize epigenetic mechanisms by utilizing *deterministic models*. These models are such that for a given input, the model outputs the same exact result. In contrast, stochastic models rely on probabilistic methods and each simulation run can give a different output depending on random decisions.

The mathematics included in this thesis are largely pedagogical meaning that the results are long known, and the emphasis is placed on succinct explanations and self-containment. In addition, a large part of this thesis is the application and interpretation of mathematical theory to biology. In an atypical manner, material on the application of mathematical concepts to biology are presented first (chapter 2, chapter 3, chapter 4). chapter 5 and chapter 6 then provide the theoretical background to the methodologies used in previous chapters. In particular, we provide brief expositions on Hidden Markov Models, Logistic Regression Models and Receiver Operating Characteristics. These chapters are accessible to anyone trained in basic calculus and probability theory.

2 A Hidden Markov Model for the discovery and identification of distinct chromatin states.

2.1 Introduction

All cells virtually share the same primary DNA sequence that encodes the genetic blueprint of an organism. Each cell-type however, has a distinct gene expression profile defined by numerous biological factors. Notably, numerous *epigenetic modifications* of *chromatin* can modulate the interpretation of the DNA sequence.

The DNA of all eukaryotic organisms is organized into the chromatin structure. This structure encodes all cellular processes such as transcription, cellular division, differentiation and DNA repair. The basic unit of chromatin is the nucleosome, a *bead like* structure that wraps 148 nucleotides of DNA and contains four core histone proteins: H2A, H2B, H3, H4 [24]. Post translational epigenetic modifications in the N-terminal tail of histones contribute the cell's specific gene expression profile and protein development. Each core histone can undergo a number of modifications such as acetylation, methylation, phosphorylation, and in multiple positions of the histone, ie mono-, di-, or tri-methylation. In particular, DNA methylation in promoters is closely associated with downstream gene expression. However, it is currently under investigation whether DNA methylation is a cause or a consequence of gene expression. Several studies suggest that DNA methylation causes changes in the affinity of transcription factors for their binding sites. Conversely, several studies suggest that that gene regulation by histone modifications is stabilized by DNA methylation.

Distinct combinatorial patterns of histone modifications play a great role in a cell's transcriptional regulatory network. More than a 100 different histone modifications have been described, leading to the so called *histone code hypothesis* that combinatorial interactions of histone modifications encodes distinct biological functions [24]. Some of these combinations are highly significant in determining cell function and morphology.

There is a growing interest in developing computational and mathematical models to capture genome-wide histone modification data. In this chapter, we identify and quantify *chromatin states*, defined to be a set of combinations of histone modifications that are biologically significant and exhibit spatio-temporal interactions [24]. A systematic genome-wide analysis is performed based on a multivariate Hidden Markov Model.

A Hidden Markov Model is a widely used statistical framework and serves many fields. The framework provides a toolkit for building complex probabilistic models and interpreting results intuitively. This power of painting a intuitive picture comes from the model's ability to label or classify underlying hidden *states* by modeling multiple observed inputs. Originally developed for computerized speech recognition, Hidden Markov Models have become paramount in computational biology. See Rabiner [48] for historical details.

We apply the methodology of Hidden Markov Models to epigenetic datasets in biology. A genome-wide analysis if performed on nine particular histone modifications based on the spatio-temporal combinations within undifferentiated and differentiated muscle cells in mouse. These epigenetic combinations are H3K18Ac, H3K9Ac, H4K12Ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, PolII. Biological significance of these modifications are already well established in literature: Ernst and Kellis [24], Asp et al. [4], and Larson et al. [41]. Given the successful biological applications of Hidden Markov Models, we conduct a genome-wide analysis of histone modifications using a Hidden Markov Model. In later chapters, we implement this model as a background model in the scoring function of the Position Weight Matrix (chapter 3) and Logistic Classifiers (chapter 4) for predicting transcription factor binding sites.

2.2 Model specification and description

Hidden Markov Models have been used successfully to model the changing landscape of DNA [24, 26, 23, 41, 8]. A number of computational and mathematical methods have been developed to systematically discover and characterize multiple epigenetic modifications. Define *chromatin states* to be distinct combinatorial patterns of epigenetic modifications (or more specifically histone modifications). We identify chromatin states on the mouse genome using on a multivariate Hidden Markov Model. In our study, we make use of the popular **ChromHMM** software, developed by Jason Ernst and Manolis Kellis, to capture combinatorial patterns and identify chromatin states [23, 24]. ChromHMM is based on a multivariate Hidden Markov

Model that models the observed combination of chromatin modifications. As input it is supplied with a multidimensional vector consisting of observed histone modifications. The software fits a Hidden Markov Model and returns the posterior probability distribution of its genome-wide state assignment. The input, in our study, was a high confidence, experimentally verified dataset of histone modifications provided by Asp et al. [4]. This dataset is generated by chromatin immunoprecipitation procedures followed by high throughput sequencing (ChIP-seq) on undifferentiated C2C12 mouse cells and differentiated cells. The ChIP-seq experiment from Asp et al. [4] yielded nine data tracks, corresponding to the raw signals of the mapped tags (or reads) of histone modifications for both cell lines. In order to systematically analyze this dataset and apply mathematical principles, the raw ChIP-seq data is processed into binarized data at a 200 nucleotide(nt) resolution. In other words, raw signals for each histone modification were converted into presence and absent values across the genome based on a Poisson background distribution. Specifically, for each histone modification, sequential intervals of length 200nt is assigned 1 if the number of reads in the interval is sufficient such that $P < 10^{-4}$ under the Poisson distribution. The mean parameter of the Poisson distribution was set to the empirical mean of mapped tags per interval. Thus, each 200nt interval has associated with it a vector of 9 boolean elements characterizing the combinatorial pattern of the chromatin modifications. This approach offers a birds-eye view of the data and reduces the chances that experimental artifact, noise, and missing data will mislead the computation. The output Hidden Markov Model, by ChromHMM, captures two types of information through its Emission Probability Matrix (EPM) and Transition Probability Matrix (TPM). The EPM captures the combinatorial patterns of the epigenetic marks and the frequency with which they occur. The TPM captures the spatial relationships of each distinct 9-length binarized vector along the genome. Under this systematic approach, genomic regions corresponding to specific functional elements such as transcription start sites, active genes, repressed genes, exons, and introns can be inferred solely from the state assigned to the region and the probability of expressing any combination of histone modifications, even though no annotation information was provided as input.

The probabilistic model We start with a fully connected topology of the underlying HMM with K states. Recall that a HMM captures the observed combinations of chromatin marks as a set of emission parameters (EPM) and models their spatial relationships with a TPM. For a state k and a histone modification m, let $p_{k,m}$ be the associated emission parameter, ie the probability that the input histone modification m has a presence call in state k. Let $v_{c_t,m}$ be the boolean value for histone modification m and interval c_t chromosome c, where t corresponds sequentially to the 200nt intervals. Denote the binary vector of HMs at interval c_t by

$v_{c_t} = [v_{c_t,1}, v_{c_t,2}, \ldots, v_{c_t,m}]$. The transition probability matrix of a HMM represents the spatial relationship of the underlying hidden states. Let $b_{i,j}$ denote the probability of transitioning from state i to state j. Let s_c be the unobserved state sequence through chromosome c, in particular let s_{c_t} be the assigned state at interval c_t. The full likelihood of the observed data, with initial probability vector a is given by

$$\Pr(\nu \mid a, b, p) = \prod_c \sum_{s_c} a_{s_{c_1}} \left(\prod_t b_{s_{c_{t-1}}, s_{c_t}} \right) \prod_t \prod_m p_{s_{c_t},m}^{v_{c_t,m}} (1 - p_{s_{c_t},m})^{1 - v_{c_t,m}} \tag{2.1}$$

The software, ChromHMM, uses a variant of the standard Baum Welch algorithm to infer the transition estimates b and emission parameter estimations p. See Online Methods in Ernst and Kellis [24] for a complete description.

Selecting a sufficient model We apply ChromHMM to the processed ChIP-seq data using the default parameters to create models of different complexities, ranging from 6 states to 18 states. The increasing complexity of a model is characterized by an increasing log-likelihood value of the models computed by the software (Figure 2.1). We selected the $K = 9$ state model for both cell types since nine states provided sufficient resolution to capture all emission parameters from higher complexity models (Figure 2.2). The lower complexity of this model allows us to resolve biologically meaningful patterns.

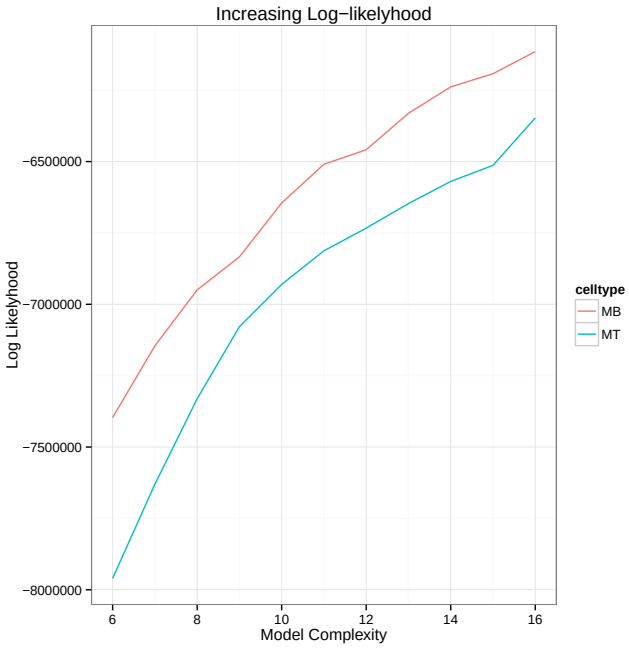


Figure 2.1: As the number of states increase in our models, it has increasing complexity, characterized by the increasing log-likelihood.

As additional validation that nine states capture fully the interactions of HMs in our dataset, consider an alphabet of $512(2^9)$ observation symbols constructed by enumerating each possible combination of modifications by mapping each 9-length HM vector to an integer value. In other words, calculating the logical `OR` of the binary values returns an integer symbol. For example, symbol *32 (0b 000 100 000)* corresponds to observing the modification H3K18Ac only, since the 4th entry in the vector is a boolean value of this modification. This approach shows that $> 95\%$ of the genome is covered by 9 dominant HM combinations (Figure 2.3). Therefore, a $K = 9$ state model is sufficient to capture the raw epigenetic information while minimizing complexity. Furthermore, the small number of states is particularly advantageous as it allows us to maximize biological interpretability. Overall, the $K = 9$ state model captures equally well the complexity of higher states models thus eliminating potentially redundant states. Mathematically, one can use Bayesian information criterion (BIC) and/or Akaike information criterion (AIC) to statistically determine the optimal model. However, in our context these methods are not effective criterion for model selection [37]. BIC and AIC favors models with more states that would be considered of biological significance [37]. However, increasing the number of states (and therefore the number of parameters to be estimated) results in an increased log-likelihood that is greater than the penalty for introducing new parameters. Thus, BIC

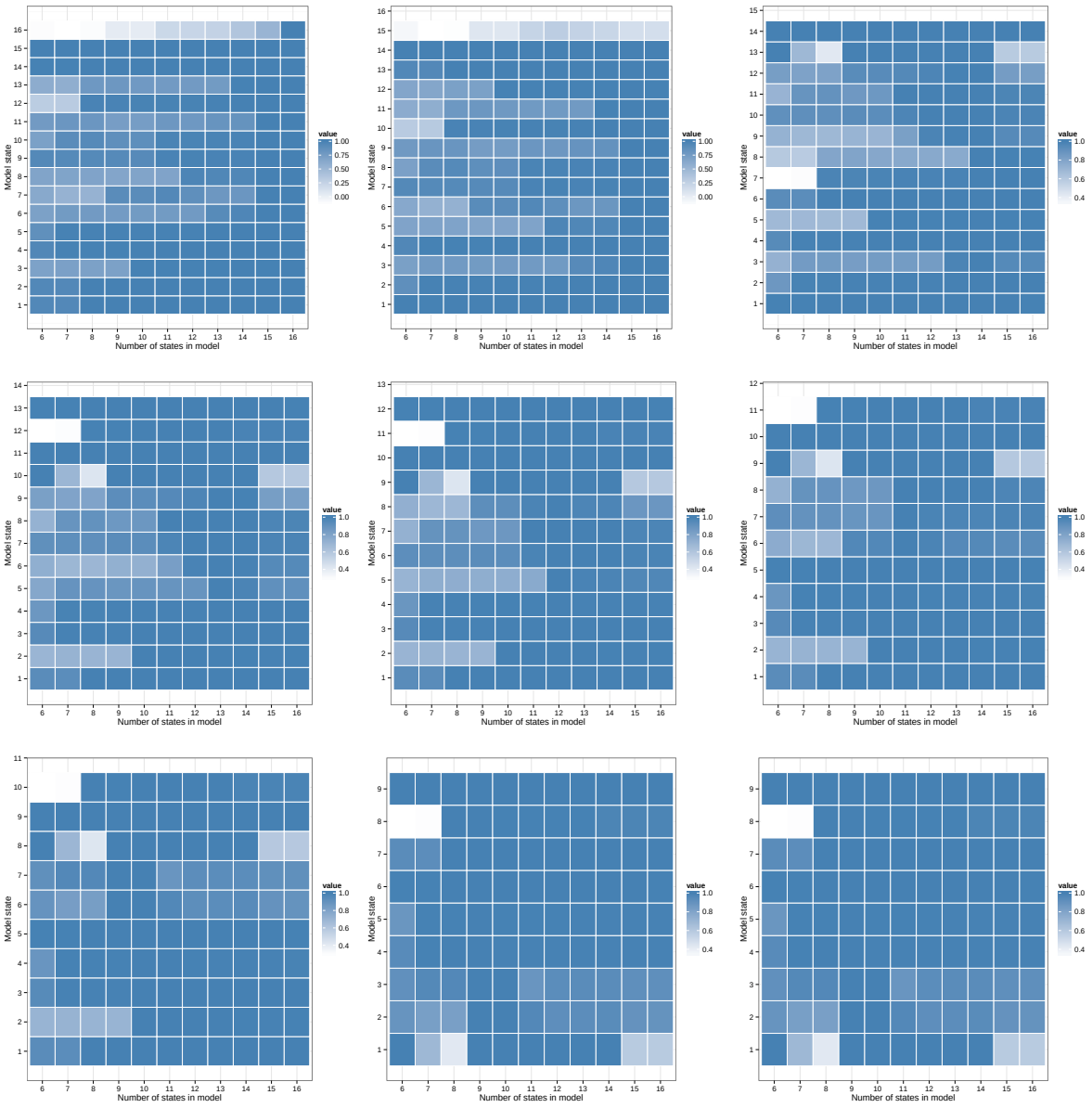
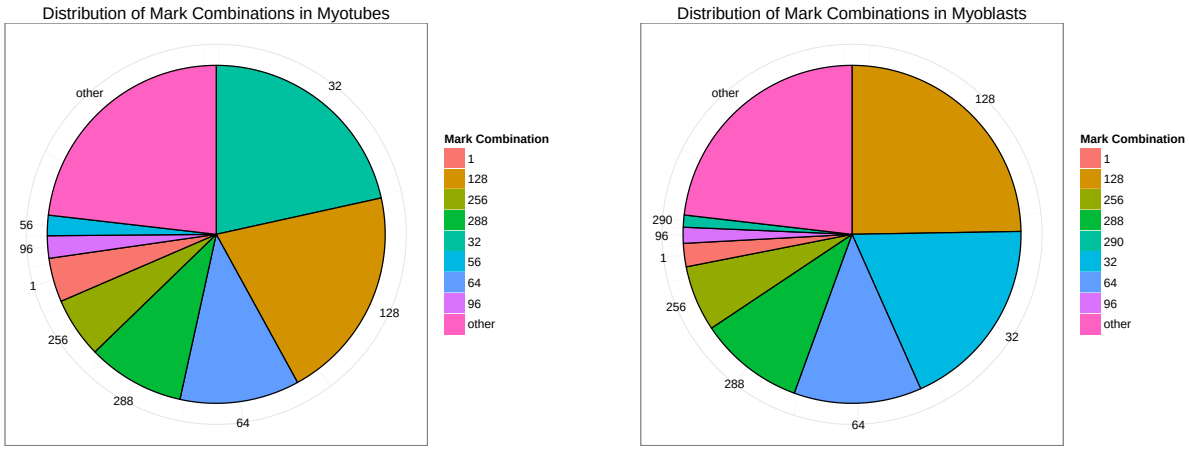


Figure 2.2: These figures compare how well our model captures the emission parameters that of higher complexity models. We see that the 9-state model has emission parameters that are highly correlated with emission parameters of higher state models. The models here are based on the MT cell type. Models on MB cell type perform return similar results.

and AIC can not help identify the most optimal and parsimonious model [24].



(a) Myoblasts cell line

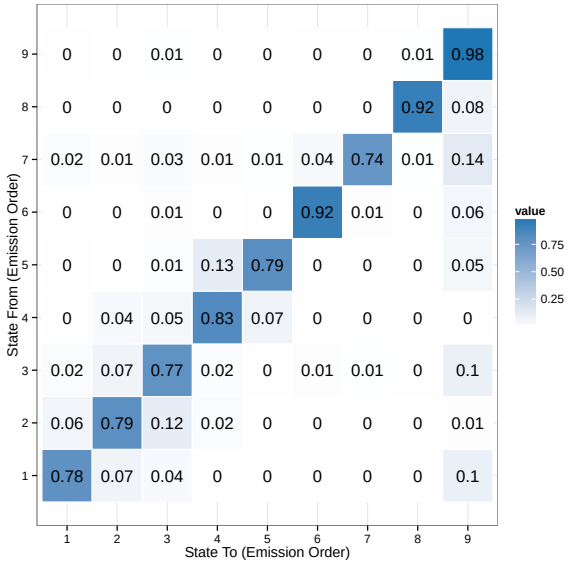
(b) Myotubes cell line

Figure 2.3: The pie charts show that the majority of the genome (95%) is covered by nine particular HM combinations. We have excluded the '0' modification (ie, no modifications at all) to emphasize the non-null combinations. 78.93% and 79.56% in both cell types, MB and MT, has no HM modifications.

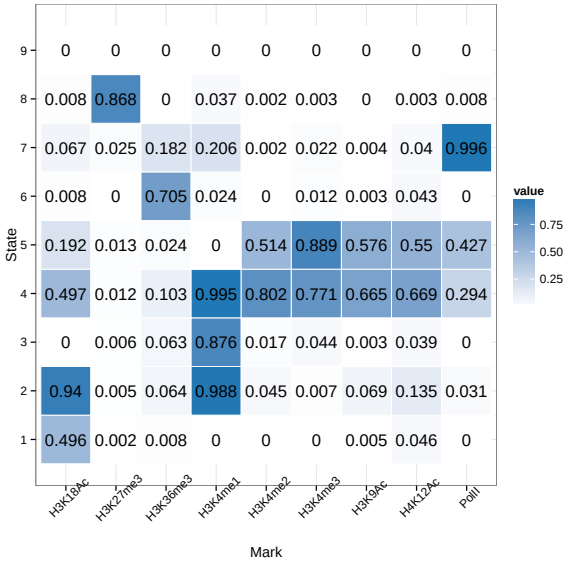
2.3 The finalized model

A Hidden Markov Model is constructed for the *de novo* identification of combinatorial epigenetic patterns in both MB and MT cell types. A 9 state Hidden Markov Model, characterized by its Transition Probability Matrix and Emission Parameter Matrix, was trained over all chromosomes where the observed sequence were combinations of histone modifications, encoded by a 9 length binary vector. The optimal state sequence of the genome was performed by the standard posterior decoding algorithm. The complete model is exhibited in Figure 2.4. The states in the model refer to the distinct combinatorial patterns of histone modifications in both cell types: myotubes and myoblasts. In other words, the 9 length emission vector associated with a given state k denotes the probability of observing each individual histone modification. The biological interpretation of the states is described below in detail.

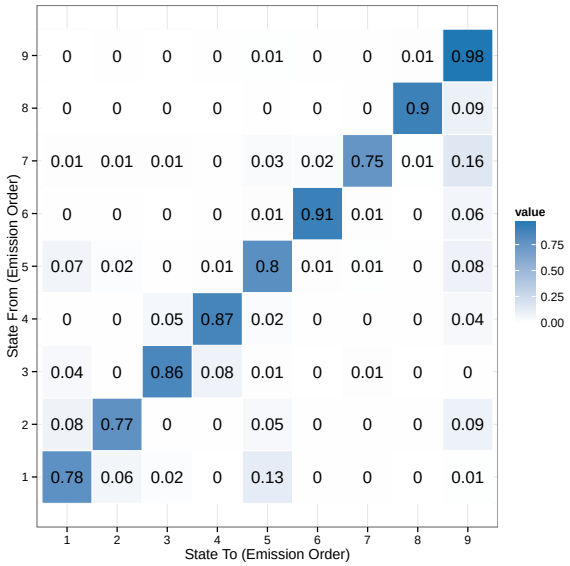
Learned transition parameters The transition probability matrix quantifies the spatial relationships between distinct chromatin states. The matrix in Figure 2.4 exhibits highly non-uniform state to state transition probabilities by having a large majority of the transition probabilities between states very small.



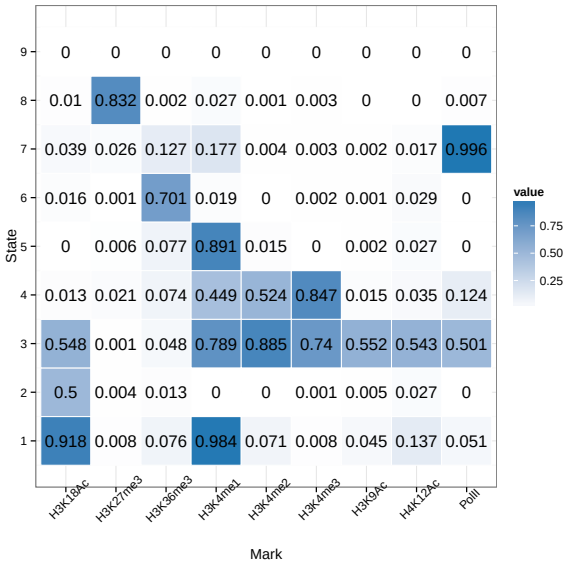
(a) TPM for Myotubes



(b) EPM for Myotubes



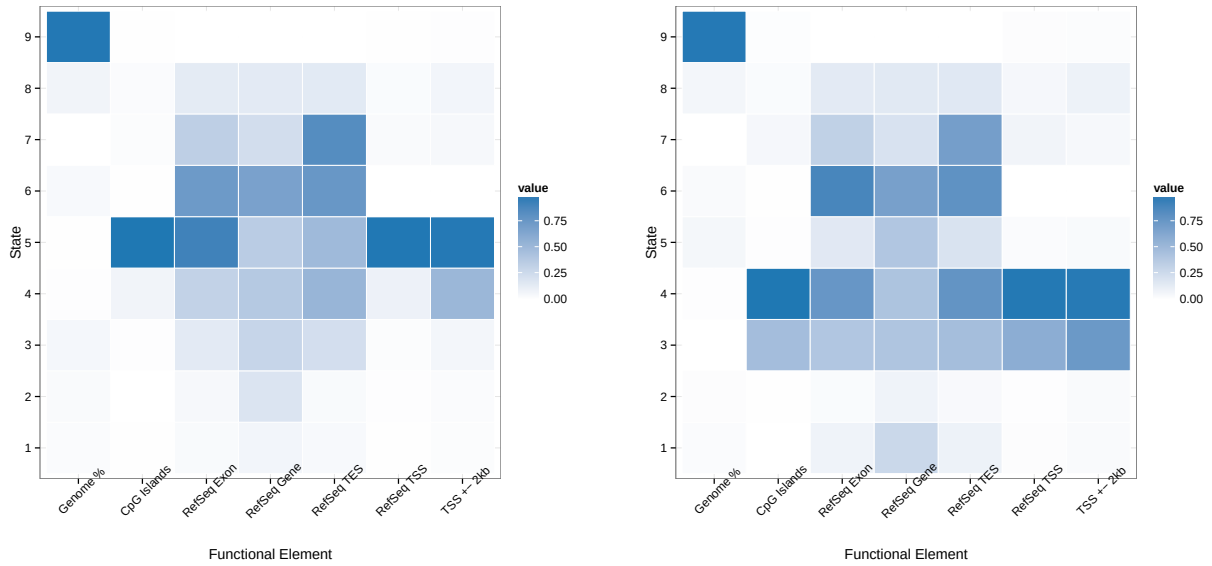
(c) TPM for Myoblasts



(d) EPM for Myoblasts

Figure 2.4: The Transition probability matrix and the Emission probability matrix for the $K = 9$ models for both cell lines as produced by ChromHMM. The transitions are from the states on the y-axis to the x-axis. Each row in the Emission Probability Matrix shows the specific combination of marks associated for the state. The color signify a value between 0 and 1 for which they occur.

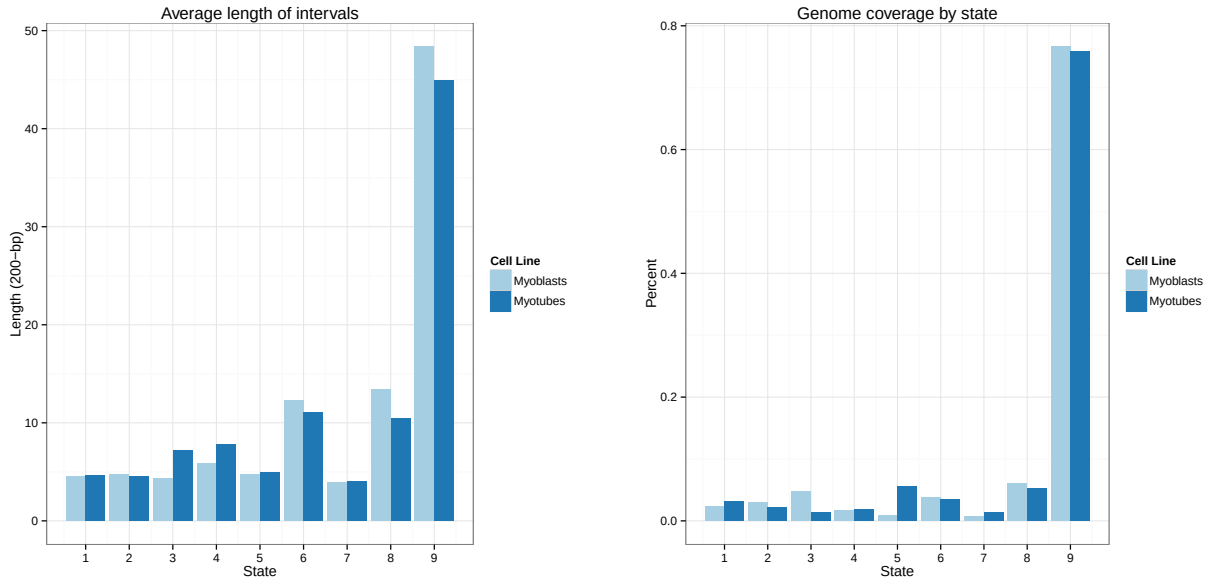
In particular, 74% of the entries in both TPMs had values ≤ 0.05. We consider the transitions that received a high probability. Active intergenic states (1 and 2) are most likely to transition to active states and to



(a) Myoblasts cell line

(b) Myotubes cell line

Figure 2.5: Enrichment of each state relative a set of external data for transcription start sites, transcription end sites, genes, exons and CpG islands. The enrichment helps identify the domain for each state.



(a) Average length of continuous state assignment.

(b) Coverage of genome by state

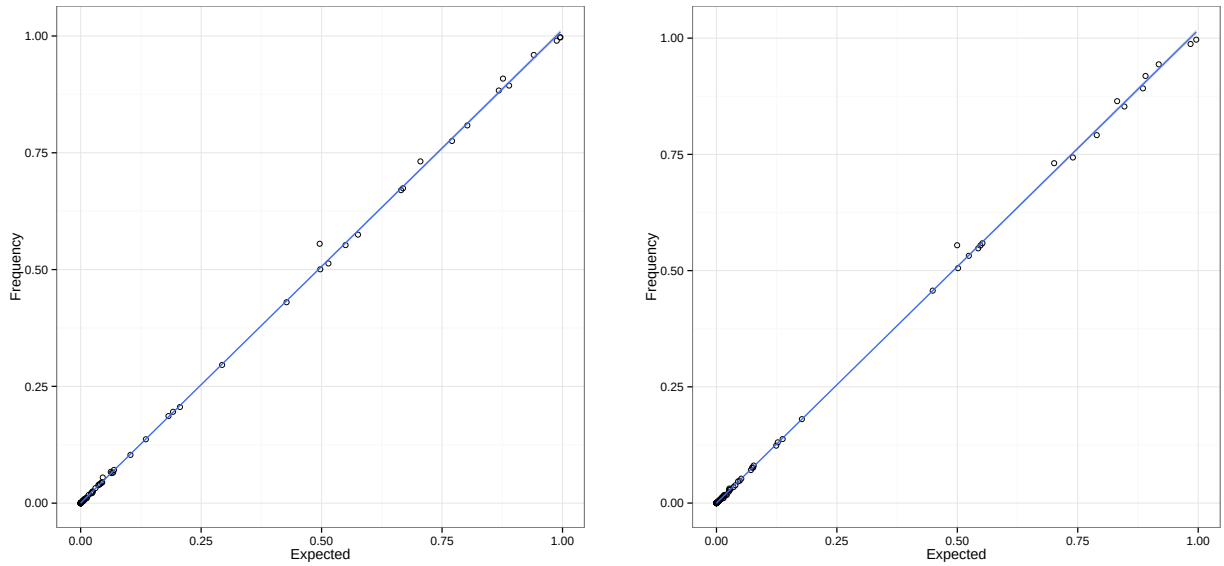
Figure 2.6: Basic statistics performed on the 9 state HMM model

further promoter states (3 - 5). Furthermore, the promoter states are highly likely to transition to other promoter states or to transcribed states. Thus, we see that the transition matrix helps define large groups of

active, promoter, transcribed states with significantly high in-group transitions than out-group transitions. This is expected as the natural progression of functional regions in the genome follows the pattern of active regions → promoter regions → transcription start site → transcribed genes. Figure 2.4 represents this by having a high diagonal in the matrices for both cell lines. The spatial relationships captured by the TPM tend to share many biological functions, validating the biological interpretability of the learned transition probabilities.

Histone modification dependency It is of importance to know how well the HMM captures the genome-wide dependencies between histone modifications. Recall that chromatin states encode combination of one or more histone modifications. First, comparing output probabilities of each HM encoded by the emission parameter to the empirical frequencies in the raw data. Figure 2.7 shows, for each fixed state, that our model is in complete agreement with the empirical data. More interestingly, chromatin states are defined by the distinct combinatorial pattern of HMs per state. If a chromatin state is defined by two or more HMs, we expect that this combination of HMs show genome-wide dependency. In other words, the particular combination of a chromatin state should occur more frequently in the raw data intervals assigned that state. In the context of our model, if the posterior decoding algorithm intervals based on the intervals' raw HM combinations into the same chromatin state, this combination become conditionally independent. Particularly, we expect the HM combination to occur within the state with the same frequency as the product of their individual probabilities. Figure 2.8 compares how often a pair of HMs is observed together (y axis) in the raw data to its expected frequency (x axis) encoded in the emission parameter. The expected frequency a pair-wise HM combination is computed by multiplying the individual emission probabilities of each HM for a fixed state. Points on the $x = y$ line are those marks for which the expected count agrees with the observed counts. Indeed, the fitted HMM shows pairs of marks occurring as expected by their individual frequencies (Figure 2.8).

Genome-wide State Discrimination The probabilistic nature of a Hidden Markov Model also offers an interpretation to the distinction of states in our model. In other words, we evaluated how distinct the 9 states in the HMM are from each other in their assignments using their posterior probabilities. The posterior probabilities of all intervals is calculated using the standard Viterbi algorithm. By analyzing the state assignment per interval, we quantified the likelihood of overlap in the genome-wide assignments



(a) Myoblasts

(b) Myotubes

Figure 2.7: Plots of the output probabilities of each mark (emission parameter probabilities) and the actual frequency of each mark in Myoblasts (A) and Myotubes (B). The blue line is the line of best fit. The perfect correlation is a line from $(0, 0)$ to $(1, 1)$. This shows that our model is in complete agreement with the observed empirical data.

of any pair of states. Particularly, we looked at the probability of a region being assigned state j given that it is assigned state i. If the state assignment of a region is not of high confidence, there is a natural expectation that different states show high probability. Figure 2.9 shows the overlap distribution of the posterior probability for all states in the HMM. Each entry (s_1, s_2) denotes the average posterior probability of being assigned state s_2 for intervals' assigned state s_1. High values off the diagonal denotes uncertainty in distinguishing between a pair of states s_1, s_2 at any fixed interval. Indeed, the strong diagonal values in Figure 2.9 shows that 9 states in our model are sufficiently distinct from each other and can be assigned different biological interpretation.

MDS Analysis The methods of Multidimensional Scaling (MDS) allows for a visual verification that learned emission parameters, capturing distinct combinations of histone modifications, are grouped together. We calculate the emission vector distances using Multi-Dimensional Scaling. In particular, we scale and project the 9-dimensional emission vectors into a 2-dimensional space. Distances are measures as 1 minutes the Pearson correlation coefficient between the vectors of emission parameters for each pair of chromatin

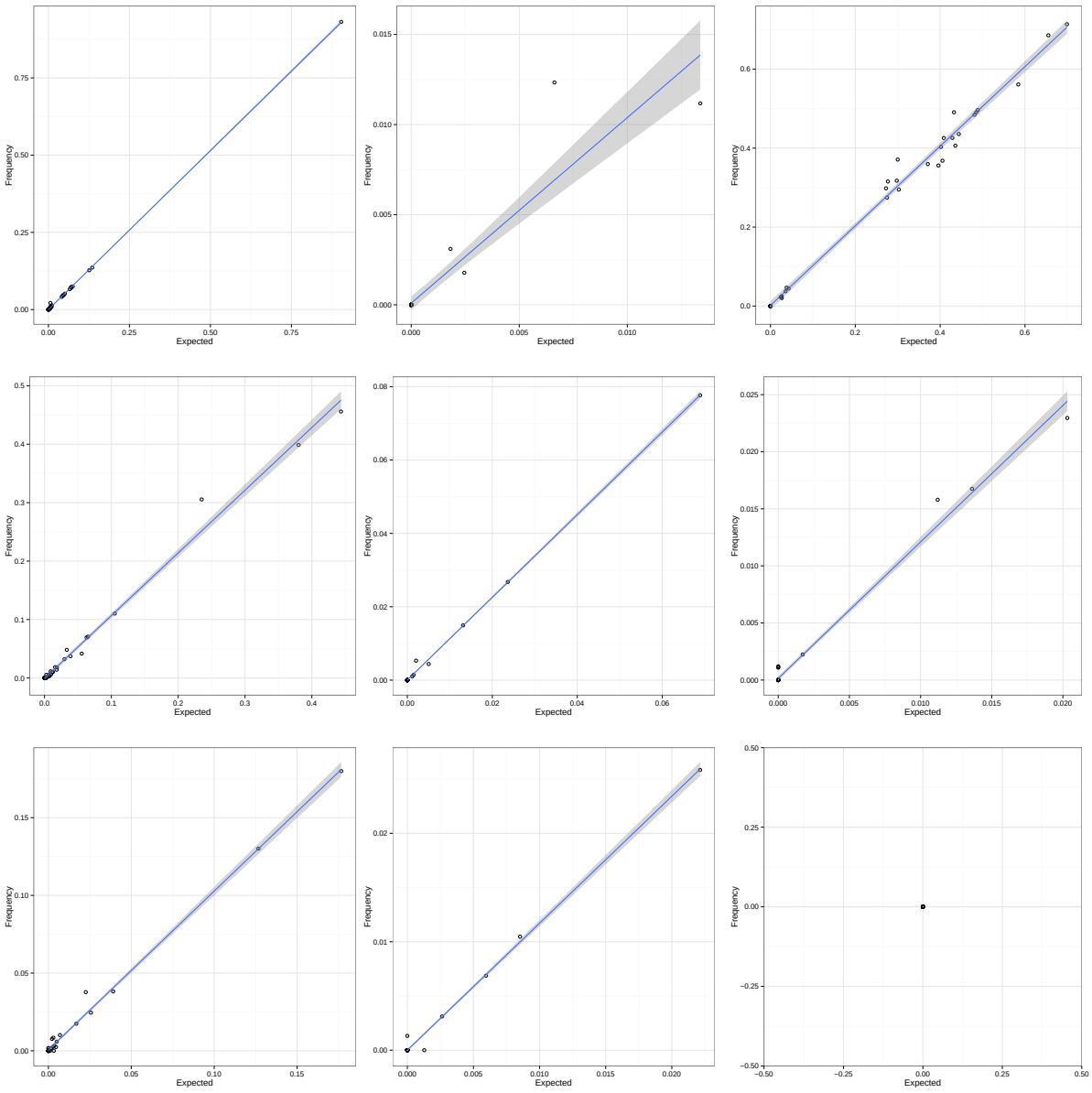
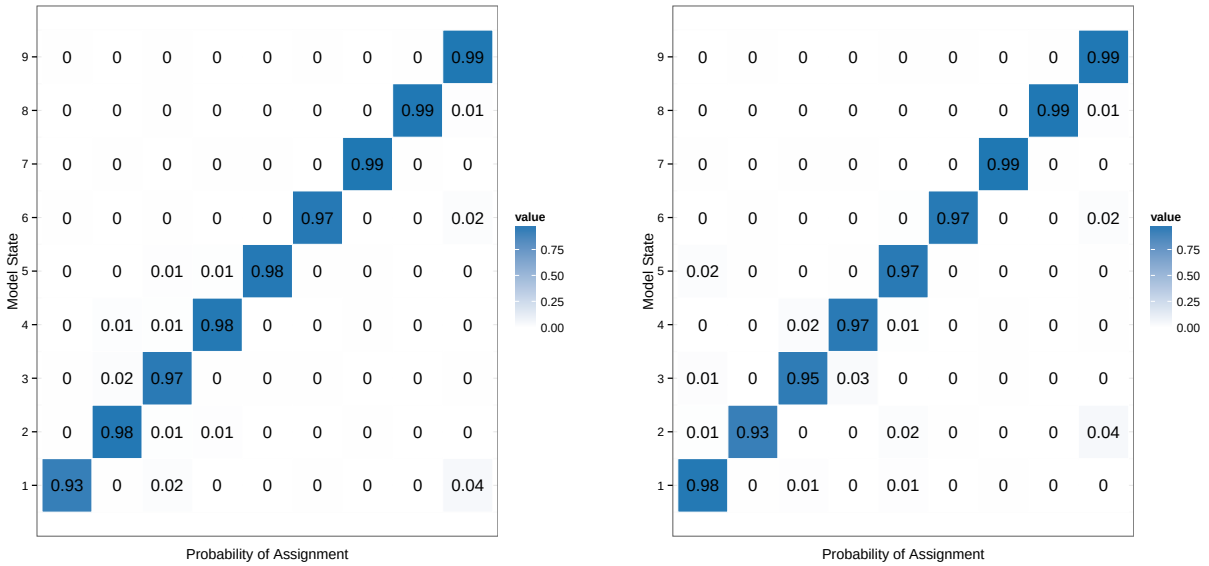


Figure 2.8: Pairwise expected vs. observed mark occurrence for each state in our 9 state model. Each plot corresponds to one state and each point corresponds to a pair of marks being observed under the model vs. how often the pair are seen in the state. The plots reveal conditional independence and validates our model assumption that conditioned on a state the pairs of marks are independent.

states. Figure 2.10 shows that the states of our model fall into distinct areas of the 2-dimensional emission space and reveals a natural grouping of the states which are consistent with the biological interpretation of each state. This is further evidence that the model's nine chromatin states capture distinct combinations of chromatin marks that cover the majority of the genome.



(a) Myoblasts cell line

(b) Myotubes cell line

Figure 2.9: Using the Viterbi algorithm, this figure summarizes the posterior probabilities for all states. In particular, each entry denotes the probability of a region being assigned state j given that its true state i. In other words, it is the frequency with which two states show probability of overlap in the same genomic interval.

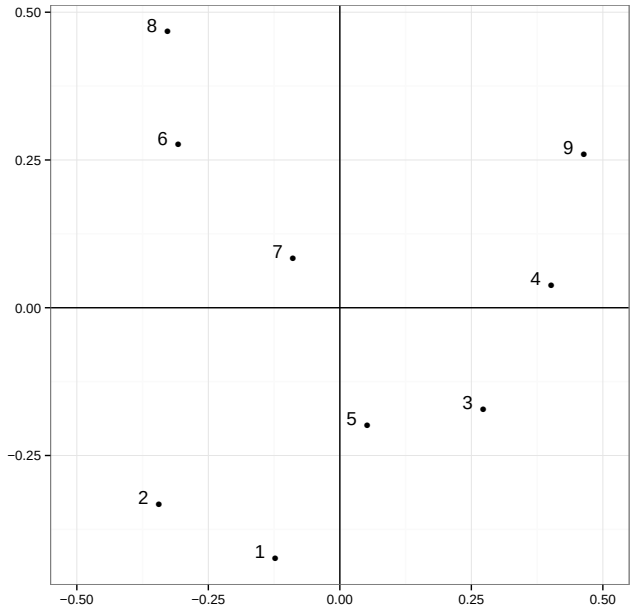


Figure 2.10: A projection of the 9-dimensional emission vectors projected into a 2-dimensional space. There exists a natural grouping of states which is largely consistent with their biological interpretation.

2.4 Interpreting chromatin states

While chromatin states are defined based on a statistical model using chromatin modification data alone, they are useful if there exists meaningful functionality and annotation of these chromatin states. The enrichment of each state of the segmentation for a set of external annotations is computed (Figure 2.5). As a result, the states vary widely in their average segment length and also exhibit varying genomic coverage (Figure 2.6). However, at this point, these states are merely an integer values and each 200bp interval is assigned this integer label. In this section, we ascertain the functional roles of these label based on a variety of evidence and investigation of histones and provide annotations for each state. Perhaps the best understood type of functional element in the human genome is the transcriptional machinery of gene expression [37]. Thus, it is reassuring that the ChromHMM's model parameters are learned in such a way so that it accounts for the identification and characterization of gene expression factors [26]. Furthermore, It is well known that chromatin plays an important role in gene regulation [4, 24, 26, 8, 41, 37]. Therefore, we expect the resulting annotation to provide diversity of genomic functions encoded by these integer chromatin states but also provide distinct differences (if any) across different cell types. To this end, we undergo a systematic integration of biological elements into the two models by assigning each integer a biological classification. From the genome wide chromatin analysis of [24, 26, 4, 41, 37], we describe the likely biological significance of the nine histone modifications in our dataset: Histone H3 lysine 4 trimethylation (H3K4me3) is modification associated with promoter regions; H3K4me2 (dimethylation) is associated with promoter and enhancer regions; The acetylations (H3K9Ac, H3K18Ac, H4K12Ac) are associated with active regulatory regions; H3K36me3 is associate with active transcribed regions; and lastly H3K27me3 is associated with Polycomb repressed regions. These modifications and their biological significance allows us to identify our states with simple summary level classifications. We annotate (classify) the nine states of our model into five general domains, emphasizing biologically meaningful differences: 1) Promoter regions and Enhancer regions including the Transcription Start Sites 2) Transcribed Regions, 3) Active regions, 4) Repressive (Polycomb repressed), and 5) Unmappable Regions. Even though the states were learned *de novo* based on the spatial relationships of histone modifications, they showed distinct association with transcription start sites, transcripts, non-coding regions, and regulatory elements.

Figure 2.5 represents the relative genome-wide enrichment in the different functional elements for both cell lines. The external data was downloaded from UCSC and included coordinates for transcription start sites,

transcription end sites, genes and exons, CpG islands, and 2000 basepairs upstream and downstream a TSS. In particular, for both cell lines, we classified states 1 and 2 as Active Regulatory states, states 3, 4, and 5 as Promoter states, states 6 and 7 as Transcribed states, state 8 as Polycomb Repressed and state 9 as unmappable or inactive states. The states exhibit variability in their continuous length with a mean length of 11.06 bins with standard deviation of 12.98 bins (Figure 2.6). The majority of the genome in both cells lines (76.6% - MB and 75.8% - MT) falls into the inactive region (state 9) which is also the large on average, with a mean length of 45 bins (9000kb). However, the active states (states 1 - 8) are smaller on average have a mean length of 6.3 bins (1.2 kb) with a standard deviation of 2.55. Thus the non-null states have less absolute variability. These properties of the model suggest that chromatin states are inherent, biologically informative feature of the genome.

Histone Modification	**Modification Type**
H2A.Z	Active
H3K4me1	Active
H3K4me2	Active
H3K4me3	Active
H3K9me1	Active
H3K9me2	Repressive
H3K9me3	Repressive
H3K27me1	Active
H3K27me2	Moderate
H3K27me3	Repressive
H3K36me1	Moderate
H3K36me3	Active
H3K79me1	Moderate
H3K79me2	Moderate
H3K79me3	Moderate
H3R2me1	Moderate
H3R2me2	Moderate
H4K20me1	Active
H4K20me3	Moderate
H4R3me2	Moderate
H2BK5me1	Active

Table 2.1: A summary of modifications grouped into active, repressive, or moderate type based on their association with active or repressed genes. Source: doi:10.1371/journal.pone.0089226.t001

Active Intergenic States The first broad class of chromatin states (states 1 and 2 for both cell lines) are classified as Intergenic states. These states, in both cell lines, had high relative frequency for H3K4me1. These states also had the highest frequency of acetylations, notably H3K18Ac. Moreover, they had low

frequencies for other methylation marks. The states were assigned to regions of genome away from the promoter regions. In fact, over 60% of the assigned states happened outside 2kb of a TSS and transcribed genes. Active intergenic states are expected to provide significant enrichments for genome-wide association study. Also, states with high frequency of H3K4me1 are associated with enhancer regions of the genome [24]. Based on the frequency of these modifications, we expect these states covered active regulatory regions of the genome such as enhancer regions, insulator regions and other regions proximal to expressed genes [24].

Promoter States In both cell lines, we classify states 3-5 as Promoter States. These states had high enrichment for promoter regions (Figure 2.5): $60 - 100\%$ of each state was within a transcription start site and $75 - 100\%$ was within 2kb of a RefSeq gene, compared with 5% genome-wide. This is further supported by their high emission probability for PolII and H3K4me2. In fact, these states all had a high frequency of methylation (mono, di, tri) of H3K4. Additionally, states 3 and 4 for myotubes and state 5 for myoblasts had relatively high CG content by having high enrichment in CpG islands (Figure 2.5) as expected of the majority of the promoters [8, 24]. The high transition probabilities to active transcribed states also support the classification of these states as promoters, especially given that probabilities are negatively affected when encountering genes on the negative strand since the promoter region comes after the gene when training the ChromHMM on the positive strand [8]. However, there exists distinct differences between the emission vectors of these states. These states differed in the frequency of other promoter-associated marks, primarily H3K4me1 and H3K4me3 and acetylations leading to varying functionality of the genomic regions assigned these states. Notably, state 4 and 5 in myoblasts and state 3 in myotubes show high frequency of acetylation (H3K9Ac, H3K18Ac, H4K12Ac). High frequency of acetylated marks often represent high expressed genes and have high enrichments for transcription factor binding [24]. The trimethlyation of the histone H3K4 along with varying levels of H3K4me1 in both cell lines suggests that these states differ in their functional promoter roles. In other words, these promoter states can be further classified into detailed descriptions such as upstream/downstream promoters, repressed promoters (high levels of H3K4me1), and transcription start sites [24, 41, 37].

Active Transcribed States Previous studies have shown that active and inactive genes are associated with different combinations of histone modifications [41]. In particular, H3K36me3 is associated with highly transcribed genes and H3K27me3 is associated with inactive genes [24, 41]. In our model, we classify states 6

and 7 for both cell lines active transcribed states. 70% of the RefSeq-annotated transcribed regions is assigned state 6 compared to 2% assignment genome-wide. Similarly, 70% of regions associated with Transcription End Sites are assigned state 7 (Figure 2.5). As additional validity, we also observed these states strongly enriched for spliced exons. As expected, the emission vector for state 6 has a high probability of H3K36me3 and the emission vector for state 7 has a high probability of PolII (Figure 2.4). Furthermore, since these states are annotated as active, the emission vectors for both states do not exhibit high probability of H3K27me3, as expected. The high enrichment of Transcription End Site for state 7 in both cell lines can be characterized by the high frequency of PolII, but also the absence of H3K4me1/2/3 often found in promoter regions. This suggests this state can be assigned a specific feature of that of the 3' ends of genes [24]. In other words, the high frequency of H3K36me3 and PolII along with a low frequency of all other modifications characterize non-promoter associated states, spliced exons, transcription end sites, and the 3' UTR regions of genes [24]. Further analysis of transcription associated modifications and their relationship with expression levels is performed in Larson *et al.* [41].

Other States State 9 for both cell lines was classified as *unmappable*. This state was assigned to a high percentage (Figure 2.6) of bins in both cell lines but exhibited very low emission probabilities for all marks. State 8 in both cell lines was classified as *repetitive and repressed* regions because of the high percentage of its bins in Repeat-Masker regions, its low emission probabilities for all marks except H3K27me3, and its very low average expression value [8, 4]. There is sufficient evidence that the histone modification H3K27me3 generated by the Polycomb repressor complex 2 (PRC2) covers repressed genes [24, 8]. Regions assigned state 8 are strong indicators that the genes within these regions have been silenced [37, 24, 8]. Furthermore, there exists a link between the histone modification H3K27me3 and myogenic differentiation. The protein complex, PRC2, required for the trimethylation of H3K27 is composed of several components including Suz12, EED, and other methyltransferases responsible [4]. In particular, removing these components, notably Suz12, accelerated myogenic differentiation and in addition cause a 2-fold increase in the number of myotubes upon terminal differentiation. This suggests that the ablation of Suz12m, and thus the loss of H3K27me3 accelerates and enhances differentiation [4]. A detailed study on the removal of the Suz12 component is found in Asp et al. [4]. Overall, in summary we classify regions assigned states 8 and 9 as heterochromatic regions, representing a large portion ($\geq$ 80%) of the genome. It may be of interest to know that in our model, we do not see states that were not expected to occur. For example, we do not expect to see high frequency of PolII and H3K27me3 occuring at the same time and no state in our model has emission vectors

that capture such a frequency [8].

Differences between MB and MT cell types There exists a natural intuition that the epigenetic landscape changes as the cell undergoes differentiation [8]. A visual inspection of Figure 2.4 suggests that majority of the combinatorial interactions of HMs in myoblasts also occur in myotubes. In particular states 1, 2, and 6 - 9 have highly correlated emission vectors, indicating that the combinatorial HM interactions encoded by these states exist in both cell types. More precisely, the model suggests there is little change in the epigenetic landscape within intergenic and transcribed regions during differentiation. However, as expected, there is a subtle difference in promoter states, ie the underlying epigenetic structure is modified in promoter regions during differentiation. In particular, the promoter states for cell type MT demonstrated higher probabilities for acetylations: H3K9Ac and H4K12Ac. Furthermore, a slight increase in PolII suggests a higher number of genes being expressed in the MT. These results are in line with observations in Bonneville and Jin [8]. The difference in states between MT and MB can be mathematically quantified. The difference score can be calculated as follows [8]

$$D(x,y) = \alpha\sqrt{\gamma\sum_{i=1}^{S}(a_{x,i}-a_{y,i})^2 + \delta\sum_{i=1}^{S}(a_{i,x}-a_{i,y})^2 - \delta(a_{x,x}-a_{y,y})^2 + \beta\sqrt{\sum_{i=1}^{N}(b_{x,i}-b_{y,i})^2}} \tag{2.2}$$

where S is the number of states, N is the number of HM combinations, $a_{x,y}$ is the probability of transition from state x to state y, $b_{x,y}$ is the emission probability of observation y of state x. The parameters α, β, δ and γ are weights for transition probabilities and emission probabilities. Borrowing the parameters from [8], $\alpha = 1, \beta = 5, \delta = 0.5, \gamma = 1$. The parameters are chosen as such due to the strong diagonal of the transition probability matrix. In other words, the parameters highlight differences of emission probabilities between states over transition probabilities between states. One may notice that formula is precisely a sum of weighted Euclidean distances, and thus state differences are symmetric.

2.5 Discussion

The general structure of chromatin and the plethora of epigenetic modifications play central roles in elucidating transcriptional machinery [24, 41]. The understanding of the epigenome is key in explaining cell development, phenotypic profiles, and disease. Improved wet-lab technologies have made generation of genome-wide histone modifications feasible. Several large projects are underway to map the interactions between histone modification. In particular, the ENCODE, modENCODE and the Epigenome Roadmap projects are global efforts to generate large amounts of HM data. Genome-wide datasets are advantageous in that their standardized nature allows for computational and mathematical methods to be easily applied.

This chapter demonstrates the power of mathematical models to provide an additional layer of genome annotation. Using a Hidden Markov Model, we identified *chromatin states* that capture distinct combinatorial patterns of epigenetic modifications in muscle cell differentiation. We find that nine distinct chromatin states capture the combinatorial interactions between the most common nine histone modifications [4]. The biological significance of each chromatin state was solely inferred based on the model's parameters. Studies show that there is a signature difference in the distribution of modifications between undifferentiated and differentiated cells [41]. Indeed, we find that there is a subtle difference in chromatin states between undifferentiated myoblasts and differentiated myotubes.

In conclusion, chromatin states offer a computational and universal way to interpret and analyze mammalian genome, especially non-protein coding regions. Most importantly, deep analysis of chromatin states can expose information about previously unannotated functional elements. This can lead to novel understanding of health and disease associated with epigenetics.

3 Quantitative Specificity of Transcription Factor Binding Sites by a Position Weight Matrix

3.1 Introduction

A significant part of cellular morphology and function is determined at the transcription level. A cell's regulation machinery underlying basal transcription consists of complex processes involving factors such chromatin modifications (chapter 2), transcription factors, RNA polymerase and other sequence specific proteins. A critical component of gene regulation relies of sequence-specific binding of multiple transcription factors to DNA sites. Thus, identifying transcription factor binding sites is key in understanding gene regulation. A variety of experimental techniques exist to determine regions bound by a transcription factor, but genome-wide binding site datasets are still rare. Current wet-lab technologies requires extensive biochemical experimentation, are costly, and time consuming. A computational approach is, therefore, inevitable and necessary.

The construction of a robust TFBS predictive model is, however, difficult and challenging because the behavior and specificity of regulatory sites is quite different that of other genomic regions. For example, restriction enzyme cleavage sites can be represented by a single DNA sequence and thus a consensus sequence model is wholly adequate. For the enzyme EcoRI, the consensus sequence is GAATTC [55] and all sites matching that pattern will be cut. Regulatory sites, in contrast, often exhibit a range of variability in bases for different sites. The consensus sequence ends up representing the 'average' sequence of the binding site. The degenerate nature of regulatory sites is biologically significant since regulatory systems can use this variability as a tool to control gene expression [55]. Stormo [54] found that, in a survey of 300 promoter regions, none of them had a binding site that was an exact match to the consensus sequence. Furthermore, this variability of sites leads to a complication that regulatory proteins (such as transcription factors) may

bind to regions (ie, have non-negligible affinity) for DNA other than their functional sites.

The simplest model consists of using the consensus sequence of the transcription factor. The consensus sequence model is simply a single DNA sequence where the base at each position is one with the highest frequency in some aligned dataset. Although, the derivation of the consensus sequence model is relatively easy, it is often not optimal in predicting sites in a random DNA sequence [55]. This problem can be alleviated by using a more general approach of using a matrix representation. The **Position Weight Matrix** introduced by geneticist Gary Stormo is an essential component in motif discovery and analysis in modern bioinformatics [54]. Elements of a PWM matrix represent the weights assigned to positions for all bases for some sequence. The *score* for any particular site is the sum of the matrix values for that corresponds to the sequence. Furthermore, note that the consensus sequence model is a special case of a PWM. Indeed, assigning a value of 1 to the element corresponding to the consequence base and 0 to all other elements yields the consensus sequence. The construction of a PWM starts with a collection of experimentally determined binding sites, in which a pattern (known as a *motif*) is extracted by aligning the sequences to maximize sequence conservation. This pattern ideally should distinguish regions of the genome that serve as binding site locations. Furthermore, the pattern, biologically speaking, is a quantitative measurement for the binding affinity of the protein. In this chapter, we provide an exposition and theoretical summary of a PWM. In addition, we derive a PWM for the *Myocyte-specific enhancer factor 2* (MEF2) transcription factor. MEF2 belongs to the MADS-box super family of regulatory protein. In vertebrates, there are four MEF2 isoforms: MEF2A, MEF2B, MEF2C, and MEF2D. It is a key transcription factor involved in the mechanics of muscle specific transcription, for both skeletal and cardiac muscle. It is also a critical protein required during embryonic and fetal development. In fact, deletion of MEF2 in embryos is fatal due to impaired heart morphogenesis [62]. Our goal is to use the MEF2-specific PWM to conduct a large scale, systematic survey to provide a more complete picture of gene regulation through MEF2.

3.2 Model specification and description

Position Weight Matrices (PWMs) are an industry standard method to represent sequence patterns also known as *motifs*. Their application is aligned with all of computational biology such that they help to elucidate regulatory mechanisms. In particular, PWMs can be used to model and provide a natural probabilistic characterization of transcription factor binding sites [54]. The PWM model is characterized by a matrix of

length w that assigns a score to a DNA sequence of length w. Sequences with high scores are expected to be candidates for potential binding sites. Figure 3.1 shows how such a model can be used to evaluate a sequence. In general, the construction of a PWM model requires three specific matrices: 1) Count Matrix 2) Frequency Matrix 3) Weight Matrix, discussed below. An important assumption in the construction of most PWMs is that the contribution from each position of the binding site are independent and additive. This simplifying assumption allows us to represent the specificity as a mono-nucleotide matrix [54, 69, 56, 55]. However, this assumption makes the score of a sequence, ie the binding affinity, an approximation for most proteins, and it remains to be seen whether it is a sufficiently good approximation [54]. A genome wide association study (GWAS) by Hoffman et al. [37] indeed shows a genome-wide functional relevance of constraints for pairs of nucleotides. In this case, a 16 row matrix where each row represents a dinucleotide would be needed to accommodate those interactions. A post-hoc sensitivity analysis determines the performance and accuracy of the PWM model.

G	**T**	**A**	**C**	**T**	**A**	**T**	**A**	**A**	**T**	**C**
			1	**2**	**3**	**4**	5	6		
		A	-28	18	1	12	10	-29		
		C	-15	-31	-12	-10	-2	-22		
		G	-18	-50	-11	-7	-11	-36		
		T	17	-17	10	-10	-5	0.49		

Table 3.1: A PWM evaluation of a sequence. Each element of the matrix corresponds to each possible base at the six positions of a DNA sequence. The matrix is used to score sliding windows of w-length subsequences. In this example, the score of the subsequence CTATAA is $= -60$.

3.2.1 Overview of the Position Weight Matrix

Let $\Sigma = \{A, T, G, C\}$, the alphabet of DNA. Let w a positive integer. A *Position Weight Matrix (PWM)* $\boldsymbol{W}$ is a function from Σ^w to $\mathbb{R}$ that assigns a number (the *score*) to each w-length *sequence* in Σ^w. Each row in $\boldsymbol{W}$ corresponds to a letter in Σ and each column in $\boldsymbol{W}$ corresponds to a position in the sequence. The matrix model calculates the score for each position along the motif by adding the relevant values in the table. That is, for each motif $u \in \Sigma^w$, the score R of u is defined to be

$$\text{Score}_{\boldsymbol{W}}(u) = R = \sum_{i=1}^{w} W(u_i, i) \tag{3.1}$$

where $u_i \in \Sigma$ is the nucleotide at the i'th position of the motif u. It has been shown that this score can be interpreted in two intuitive ways. The first is using a thermodynamics approach in which the score is an estimate of the free energy of the protein binding to this sequence site. The second approach is a statistical one where the significance of the results is dependent on the sample size. In this approach we look at the log likelihood ratio for the hypothesis that a potential binding site is found under the frequency model vs the hypothesis that a potential binding site is found under the background model. This is discussed in detail in section 3.4.

3.2.2 Determining the elements of M

The Position Weight Matrix generally involves working with a count matrix, a frequency matrix, and a log matrix. To determine the elements of these matrices, we use a collection of high confidence, experimentally verified, aligned binding sites, (see section 3.3 for details). The *count matrix* is determined by counting the number of bases in each position of every site in this collection. Denote the elements of the count matrix as $n(b, i)$ where (b, i) refer to the base and position (column) of the matrix. The *frequency matrix* is the frequency of bases at each position, where each entry is derived from the count matrix

$$n_f(b, i) = n(b, i)/N$$

where N is the total number of binding sites in the collection of known binding sites. Since the model is constructed with a finite number of sequences exhibiting variability, a nucleotide $\in \Sigma$ need not occur at least once in a particular position. In other words, if N is small, a nucleotide $b \in \Sigma$ may not be observed at a particular position i, thus having a count of zero (or too small a value). This imposes a harsh penalty and can sway our beliefs from the neutral hypothesis that all nucleotides contribute independently and equally. It is, therefore, a common practice to include a *smoothing* parameter, often referred as *pseudocounts*, added to frequency $n_f(b, i)$ values [46, 30, 54] . Pseudocounts can be constant value, proportional to a nucleotide's background frequency, or inferred from the information already gathered on the nucleotide signal. Mathematically, pseudocounts are motivated by Bayesian statistics. In biological datasets, it is common to assume a Dirichlet prior distribution for nucleotide frequencies, so that the mean estimator is equivalent to adding pseudocounts to the observed counts. If we consider a simple case where the pseudocounts are inferred based on their background distribution, a suitable expression for the pseudocount added frequencies

is

$$F(b,i) = \frac{n(b,i) + s(b)}{N + \sum_{b'} s(b')} \tag{3.2}$$

where $s(b)$ denotes the pseudocount function for base $b \in \Sigma$. Often, $s(b) = 0.25 \, \forall \, b \in \Sigma$. We provide a brief exposition on deriving $s(b)$ in Appendix A. A more detailed and theoretical study on optimal pseudocounts can be found in Claverie and Audic [10] and Nishida, Frith, and Nakai [46].

Elements of the the PWM $\boldsymbol{W}$ are derived using a combination of thermodynamical and statistical likelihood principles. The following section provides a brief exposition on constructing a PWM model based on the log-odds scores of the observed frequencies of each base compared to the background frequencies.

Thermodynamical and Statistical Methods The interaction between a transcription factor and a particular DNA sequence, u, is governed by the reaction association rate k_{on} and the dissociation rate k_{off} for the formation of the protein-DNA complex [56]. The equilibrium binding constant of the transcription factor is

$$K_{eq} = \frac{[TF \cdot u]}{[TF][u]}$$

A convenient way to quantify the *specificity* of transcription factors is to normalize K_{eq} to some reference value defined by the user [56]. The dissociation constant k_{off} follows the relationship $1/k_{on}$. The molar *Gibbs free energy* (the binding affinity) ΔG is then related to the dissociation constant k_{off} by

$$\Delta G = RT \ln k_{off}$$

where R, T are the ideal gas constant and temperature [56, 17]. In a simple experiment where only a single sequence u is available for binding, u can be in two possible states: bound or free, indicated by a binary variable $B = 1$ or $B = 0$ respectively. The probability of the sequence u bound by a transcription factor is given by

$$\Pr(B = 1 \mid u) = \frac{[TF \cdot u]}{[TF \cdot u] + [u]} = \frac{1}{1 + \frac{1}{K_{eq}[TF]}} = \frac{1}{1 + e^{E(u) - \mu}} \tag{3.3}$$

where $E(u) = -\ln K_{eq}$ is the standard free energy of binding to sequence u, and $\mu = \ln[TF]$ is the chemical potential [17, 69]. The probability can be interpreted that a sequence with binding energy below the chemical potential is almost always bound to a protein. The binding energy $E(u)$ can be decomposed into two modes: non-specific binding that is independent of the sequence and specific binding that depends on the sequence

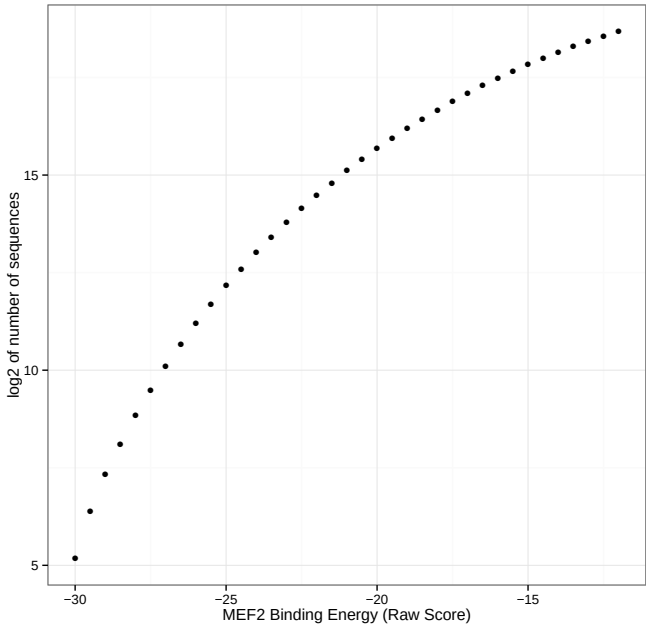


Figure 3.1: Binding probability as a function of binding energy, by

u [69], ie

$$e^{-E(u)} = e^{-E_{sp}(u)} + e^{-E_{ns}} \tag{3.4}$$

We focus our attention to the specific binding energy $E_{sp}(u)$ which is a function of the sequence u. The assumption of additivity that each base contributes independently to the total energy allows this function to be represented as a position weight matrix [69]. In other words, the individual energy contribution by base b at position i, denoted $\epsilon(b, i)$ models the energy function $E_{sp}(u)$ as

$$E_{sp}(u) = \sum_{b \in \Sigma} \sum_{i=1}^{w} u(b, i)\epsilon(b, i) = \boldsymbol{\epsilon} \cdot \boldsymbol{u} \tag{3.5}$$

where $u(b, i)$ is an indicator variable such that $u(b, i) = 1$ if nucleotide b occurs at the i'th position of sequence u. The existence of $E(u)$ is guaranteed; in the worst case, we can define a priori a list of binding energies to all possible sequences $\in \Sigma^w$ so that $E(u)$ returns the the value of $u \in \Sigma^w$ on this list.

Equation 3.3 holds true in the simple case where a single sequence u is available for binding [69, 36], however it is also true for the general case where many different sequences (and in different proportions), are competing for the same transcription factor which exists with some known concentration. [69, 36]. Suppose the binding

site for the transcription factor of interest has a fixed length of w nucleotides; thus there are 4^w different sequences to which the transcription factor could bind. Recall, the set of all 4^w possible binding sites was defined as Σ^w. For the sake of clarity, we make this rigorous:

$$\Sigma^w = \{u_i \mid 1 \leq i \leq 4^w \text{ and each } u_i \text{ is of length } w\}$$

where $\Sigma = \{A, C, G, T\}$. Consider an experiment where a single transcription factor is present alongside all $u \in \Sigma^w$, with each u_i (i is indexed over the set Σ^w) occurring with proportion $\pi(u_i)$. In this situation, the transcription factor not bound to u_i could be bound to some other sequence $u_j, i \neq j$. Recall that μ represents the chemical potential. In this context, however, μ corresponds to the average free energy for the collection of sequences not bound by the TF. A sequence $u_i \in \Sigma^w$, at any particular moment, can be in three states: bound to the TF by specific binding, bound to the TF by non-specific binding, and not bound at all. The probability of each state is determined by the energy of that state, and when the concentration of the factor is low that all sites have very low probability of being bound, the probability of each state is governed by the Boltzmann distribution [36, 69, 17]

$$\begin{aligned}
\Pr_{sp}(B = 1 \mid u_i) &= \frac{e^{-E_{sp}(u_i)}}{e^{-\mu} + e^{-E_{sp}(u_i)} + e^{-E_{ns}}} \\
\Pr_{ns}(B = 1 \mid u_i) &= \frac{e^{-E_{ns}}}{e^{-\mu} + e^{-E_{sp}(u_i)} + e^{-E_{ns}}} \\
\Pr(B = 0 \mid u_i) &= \frac{e^{\mu}}{e^{-\mu} + e^{-E_{sp}(u_i)} + e^{-E_{ns}}}
\end{aligned}$$

The overall probability of a sequence u_i being bound is, therefore, the sum of the above probabilities, thus

$$\begin{aligned}
\Pr(B = 1 \mid u_i) &= \Pr_{sp}(B = 1 \mid u_i) + \Pr_{ns}(B = 1 \mid u_i) \\
\Pr(B = 1 \mid u_i) &= \frac{e^{-E(u_i)}}{e^{-\mu} + e^{-E(u_i)}} \qquad (3.6)
\end{aligned}$$

which is equivalent to Equation 3.3. Applying Bayes' Rule to (3.6) gives the probability of all bound sequences out of all sequences in Σ^w:

$$\Pr(u_i \mid B = 1) = \frac{\frac{e^{-E(u_i)}}{e^{-\mu}+e^{-E(u_i)}}\pi(u_i)}{\sum_j \frac{e^{-E(u_j)}}{e^{-\mu}+e^{-E(u_j)}}\pi(u_j)} = \frac{\frac{e^{-E(u_i)}}{e^{-\mu}+e^{-E(u_i)}}\pi(u_i)}{Z} \qquad (3.7)$$

where $Z = \sum_j \frac{e^{-E(u_j)}}{e^{-\mu}+e^{-E(u_j)}}\pi(u_j)$ is the so called *partition function* required so that $\sum_i \Pr(u_i \mid B = 1) = 1$. In our definitions, we have not specified the temperature or the ideal gas constant, both very important factors in reaction kinetics. However, the above derivations could easily be applied with the replacement of $e^{-E(\cdot)}$ with $e^{\frac{-E(\cdot)}{RT}}$. We could also modify our function $E(\cdot)$ by adding a constant $E'(\cdot) = E(\cdot) + c$. If we choose $c = \ln Z$, then $\sum_u \pi(u)e^{-E'(u)} = 1$. This has an important implication: that is we are able to choose our baseline of energy so that the probability of of any particular site being bound is simply the negative logarithm, ie

$$P(u \mid B = 1) = e^{-E(u)}\pi(u)$$

Equations 3.7 and 3.5 provide a complete description of the PWM model. Substituting (3.5) into (3.4), and then into (3.7), we obtain the relationship between the statistical probability of a bound sequence u and its thermodynamical binding energy $E(u)$. The unknown parameters $\theta = \{\boldsymbol{\epsilon}, \mu, E_{ns}\}$ are estimated, and in particular we are interested in the parameter $\boldsymbol{\epsilon}$ used to construct the model.

Estimation by Maximum Likelihood methods Let $F(u) = \Pr(u \mid B = 1)$. F can be interpreted as the fraction of time for which a sequence u_i will be bound [36]. Alternatively, from (3.7), $F(\cdot)$ corresponds to a value that is directly proportional to the sequence's binding affinity, given by equation 3.5 [17]. Given a large enough sample of bound sequences, these probabilities can be used to estimate the energy function $E(u)$ by maximizing $F(\cdot)$. Furthermore, since $\pi(u) \geq 0$ for all $u \in \Sigma$, it is of interest to know the fraction of time a transcription factor binds to any particular u out of its copies. This is biologically intuitive since transcription factor can bind specifically or non-specifically. However, only a fraction of bound sequences are involved in gene regulation. The ratio $F = e^{-E}\pi$ determines the amount of binding to a particular site relative to the background of all possible sites.

The unknown parameters can be estimated by well established methods such as Maximum Likelihood Estimation [69, 17, 55], Bayesian Statistics, or Machine Learning [36]. As pointed out by Djordjevic, Sengupta, and Shraiman [17], the log-odds method ((3.7)) is only applicable in a special case of (3.6). In particular, it is only applicable when the concentration of the transcription factor is low ($\mu \to -\infty$). Therefore, as suggested by Djordjevic, Sengupta, and Shraiman [17], (3.6) can be replaced, ie

$$\Pr(B = 1 \mid u) = e^{-E_i}e^{-\mu} = e^{-H_i}$$

To derive the objective likelihood function, consider an experiment in which a large number of sequences of length w are generated and made available to a transcription factor with a known concentration. Let $\pi(u)$ denote the the proportion (or probability) of a sequence u in this experiment. At equilibrium, the transcription factor is extracted, along with the bound DNA and sequenced. This gives us a set O containing n_O sequences that are all bound by the transcription factor. The likelihood function is derived by considering the binding energy for all sites in O relative to the background of all possible sites. In other words, we need to maximize $F(u) = \frac{e^{-H_i}\pi(u)}{Z}$ where Z is the partition function. Thus, the likelihood function of observing the sequences $\in$ O to be maximized is given by

$$e^{\pounds} = \prod_{S \in \mathrm{O}} Z\pi(S)e^{-H(S)} \tag{3.8}$$

or, instead, maximizing the log-likelihood

$$\pounds = n_O \ln(Z) + \sum_{S \in \mathrm{O}} \ln(\pi(s)e^{-H(S)}) \tag{3.9}$$

Note that the partition function Z creates a complication for maximizing the likelihood function. The issue is that for larger values of w, the calculation of Z by the naive approach of enumerating over all sequences becomes computationally infeasible. However, if one assumes a random genome, the additivity assumption that each position contributes independently to the total binding energy allows Z to be derived analytically [36]. Although, genomes are not random sequences, short subsequences occur with frequencies according to a uniform background model. Therefore, the proportion π for a sequence $u \in \Sigma^w$ of length w can be computed assuming independent, identically distributed bases with composition $\pi_{bg}(b)$

$$\pi(u) = \prod_m \prod_b \pi_{bg}(b)^{u(b,m)}$$

where $u(b, m)$ acts as a selector such that only one value of $\pi_{bg}(b)$ is used in the product for each position m. Thus

$$\pi_{bg}e^{-H(u)} = \prod_m \prod_b \left(\pi_{bg}(b)e^{-H(b,m)}\right)^{u(b,m)}$$

Summing over all sequences $u \in \Sigma$ computes the partition function Z. Recall that we are interested in the maximization problem which solves for the unknown parameter $\theta = \{\epsilon(b, i)\}$. A detailed step-by-step solution to solving the maximization problem is outline in the supplementary files of Djordjevic, Sengupta,

and Shraiman [17]. The function ϵ that maximizes the probability of binding to the collection of known binding sites is given by

$$\epsilon(b,k) = \ln\left(\frac{F(b,k)}{\pi_{bg}(b)}\right) \tag{3.10}$$

where $F_{b,i}$ is the pseudocount added frequencies, $\pi_{bg}(b)$ represents the background frequencies of each base in the genome. The matrix function $\boldsymbol{\epsilon}$ is well established in literature, and is the so called **position weight matrix** [54, 55, 36, 29, 69, 38]. Heumann, Lapedes, and Stormo [36] reaches (3.10) by applying machine learning methods. In their study, using the underlying assumption that each base position contributes linearly and independently, the perceptron neural network tries to maximize (3.9). Coupled with the analytically derived partition function, the neural network solves the maximization problem and returns exactly (3.10) for the weight matrix $W(b,k)$. Similarly, Djordjevic, Sengupta, and Shraiman [17] approaches the maximization problem using using Quadratic Programming algorithms, and similar results are found. Zhao, Granas, and Stormo [69] gets to the solution by using a model $N_i = \hat{N}_i + err$ for predicted number of binding site occurrences, where err followed a zero mean Gaussian. In this context, the probability of the data, with parameters θ is

$$\Pr(\text{data} \mid \theta) = \prod_j \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} e^{\frac{(\hat{N}_i - N_i)^2}{2\sigma^2}}\right)$$

Their study uses the Levenberg-Marquardt algorithm to solve the non-linear parameter estimation problem.

In this thesis, we define a new PWM matrix $\boldsymbol{W}$ that is slightly modified. The elements of the $\boldsymbol{W}$ are similar to the weight matrix $\boldsymbol{\epsilon}$ however, we arbitrarily take *base two* log instead of the natural log to measure the information content of the model in bits. Therefore, in conclusion, the final elements of the PWM are

$$W(b,i) = \log_2\left(\frac{F(b,i)}{\pi_{bg}(b)}\right) \tag{3.11}$$

Recall that by the thermodynamical approach, the entries of $\boldsymbol{W}$ are maximum estimates for the binding energy contribution of each base at each position of a sequence. In pure statistical sense, however, the entries of $\mathbf{W}$ normalizes the frequencies of bases in our model's training set to the *a priori* frequencies of obtaining each base [54]. This allows us to define the **Information Content** of a Position Weight Matrix: A measure of discrimination between different sites bound by the transcription factor [54, 55].

Information Content Nucleic acid data is often modeled by a 0-order Markov Chain. In this model, the four DNA bases are distributed identically and independently thus making the letters of a sequence independent and identically distributed [35]. Therefore, the probability of a particular sequence is the product of the probability of the individual letters. The probability of the individual letters **(the background model)** is the frequency of letters in an organism's genome. Most PWM analysis software such as FIMO [32] use the entire genome as the background model. The most interesting PWM is one such that the letter frequencies per column most differ from the background model [35, 56, 17]. The log-likelihood ratio is a measure used to characterize this divergence [35] and is defined as

$$\text{log-likelyhood ratio} = \sum_{j=1}^{w} \sum_{b \in \Sigma} n(b,i) \log_2 \frac{\pi_{bg}(b)}{F(b,i)}$$

The **Information Content** I_{seq} is a statistic of a position weight matrix and is the normalized loglikelyhood ratio. This statistic is widely used to estimate the statistical significance of the results of Position Weight Matrix [55, 35]. The information content of a column i can be mathematically described as

$$\text{IC}_{\text{seq}}(i) = \sum_{b \in \Sigma} F(b,i) \log_2 \frac{F(b,i)}{\pi_{bg}(b)} \tag{3.12}$$

where $F(b,i)$ is the pseudocount added frequencies of bases in the aligned dataset. Readers may notice that this is the dot product of the frequency matrix $\boldsymbol{F}$ and the PWM matrix $\boldsymbol{W}$. I_{seq} represents the Kullback-Liebler divergence or **relative entropy** [35]. The information content per column is a measure of how conserved the particular base is at that position. The maximum measure at any position is 2 bits which corresponds to only one base being allowed at that position. I_{seq} is also related to the thermodynamics of biology. Recall that the elements of the matrix $\boldsymbol{W}$ are such that they maximize the probability of binding to the collection of known function sites used to generate the PWM. The information content, then, is the relationship between the average ΔG of the protein binding a functional DNA site and the ΔG of the protein binding an arbitrary DNA sequence [55, 54, 35, 17, 56]. In other words, the I_{seq} is a measure of the difference between the probability distribution of the Position Weight Matrix and the uniform distribution. The sum of the the information content per column ($\sum_{i=1}^{w} I_{\text{seq}}(i)$) is a measure of the distance from the center of the distribution where $F(b,i) = \pi(b)$ [35]. In general, nucleic acids PWMs tend to have a lower information content than in proteins [10].

3.3 A Model for Myocyte Enhancer Factor 2

Myocyte Enhancer Factor 2 Myocyte Enhancer factor-2 (MEF2) is a transcription factor involved in the regulation of cardiac and skeletal muscle genes. It is a member of the MADS(MCM1, agamous, deficiens, serum response factor)-box transcription factors, and plays a profound role in muscle cells to control myogenesis and morphogenesis [62]. MEF2 proteins act synergistically with other transcription factors (protein-protein interactions), in particular the MyoD family, to regulate a certain set of target muscle genes [62, 29]. The transcription factor binds to a highly conserved DNA sequence in the control regions of muscle-specific genes [7]. Furthermore, MEF2 is an essential component in gene regulation of embryonic and fetal development as well as post-natal gene regulation for tissue homeostasis [62]. In fact, loss of MEF2 during early stages of cell differentiation is fatal due to impaired heart morphogenesis [62]. Given the fundamental role of MEF2 in muscle differentiation, discovery of its binding sites will further elucidate regulatory machinery. In vertebrates, there are four isoforms of the MEF2 gene (A-D), that all bind to the consensus sequence (C/T TA $(A/T)_4$ TA G/A) [62, 7]. Discovering binding sites by a consensus sequence model tends to have poor accuracy (need more citations) [54, 29]. However sufficient information has been collected thus far to enable a Position Weight Matrix model for binding site discovery [29, 28, 62]. In a previous study by Fickett [29], it has been shown that a PWM model allows discrimination of naturally occurring MEF2 sites with high sensitivity and specificity. Improving the accuracy, however, has been difficult by the fact that MEF2 combinatorially interactions with other transcription factors. In the following sections, we use a collection of experimentally verified binding sites as a training set to construct a Position Weight Matrix. In addition, we perform sensitivity and specificity analysis at a small scale as well as large scale.

Selection data Constructing a PWM requires (i) an existing motif consensus (ii) a list of experimentally verified binding sites and (iii) a database of sequences expected to be enriched in the TFBS of interest and a control set. The genome-wide human set of high confidence predicted binding sites for the MEF2 family of transcription factors were selected from the FANTOM 4 database (`http://fantom.gsc.riken.jp/4/download/GenomeBrowser/hg18/TFBS_CAGE/allsites_cage_tfbs_feb09_latest.gff.gz`). The database FANTOM 4 is an international effort to annotating and describing the regulatory mechanisms of mammalian cells. The above link downloads the *gff* file corresponding to the binding sites in the human reference genome (NCBI Build 36.1, "hg18"). Neither the alignment nor the nucleotide frequencies at positions within the

sites were known. This positional information is required to construct the position weight matrix, and therefore we used the multiple alignment software *MAFFT* to align the binding sites. This resulted in a block alignment of $N = 1875$ binding sites.

Constructing the model The Count Matrix and the Frequency Matrix, shown in Tables 3.2a and 3.2b are constructed using the block alignment of $N = 1875$ sites. A pseudocount, calculated using the *constant mode* (see Appendix) of $s(b) = 0.25$ for all bases $b \in \Sigma$. Since the modified frequency $F(b, i)$ satisfies the property

$$F(b, i) \to n_f(b, i) = \frac{n(b, i)}{N}$$

the smoothing effect of the pseudocount is negligible. The elements of the final PWM matrix $\boldsymbol{W}$ (Table 3.2c) are derived using the expression in Equation (3.11). In deriving the log matrix, the background probability distribution for nucleotide frequency

$$\pi_{bg}(b) = \{A = 0.291, C = 0.208, G = 0.208, T = 0.291\}$$

, the background frequencies of NCBI Build 37 ("**mm9**") of the *Mus Musculus* genome, primarily because all our further analysis is conducted on the mouse genome. As a control, we utilize a widely used motif finding tool to discover a PWM model in our N-wide block alignment selection data. We use the *MEME* software from the Motif based sequence analysis toolkit, *MEME Suite* [6]. The software returned the count matrix in Table 3.3. To compare the two matrices, we evaluate a score $0 \leq s \leq 1$. This score is a normalized version of the sum of column correlations as proposed by Pietrokovski (1996). This score 0.9156568 suggests that the PWM in Table 3.2 and the PWM obtained from external software are similar. This further adds validity to the thermodynamical derivations in the preceding section.

3.4 Distribution of the scores

The ultimate goal of a PWM model is to be able to discover novel binding sites. In the absence of experimentally verified binding site locations, the expected rate of false positives can be computed by considering the statistical significance of scores. In this section, we formulate two characterizations of *statistical significance* of a PWM model. First, we would like to know how independent positions of the PWM contribute

	1	2	3	4	5	6	7	8	9	10
A	518	91	1515	773	1054	880	1276	637	1272	502
C	809	631	72	107	24	31	73	63	18	117
G	84	30	121	86	33	37	25	96	524	1130
T	464	1123	167	909	764	927	501	1079	61	126

(a) Count Matrix

	1	2	3	4	5	6	7	8	9	10
A	0.28	0.05	0.81	0.41	0.56	0.47	0.68	0.34	0.68	0.27
C	0.43	0.34	0.04	0.06	0.01	0.02	0.04	0.03	0.01	0.06
G	0.04	0.02	0.06	0.05	0.02	0.02	0.01	0.05	0.28	0.60
T	0.25	0.60	0.09	0.48	0.41	0.49	0.27	0.58	0.03	0.07

(b) Frequency Matrix

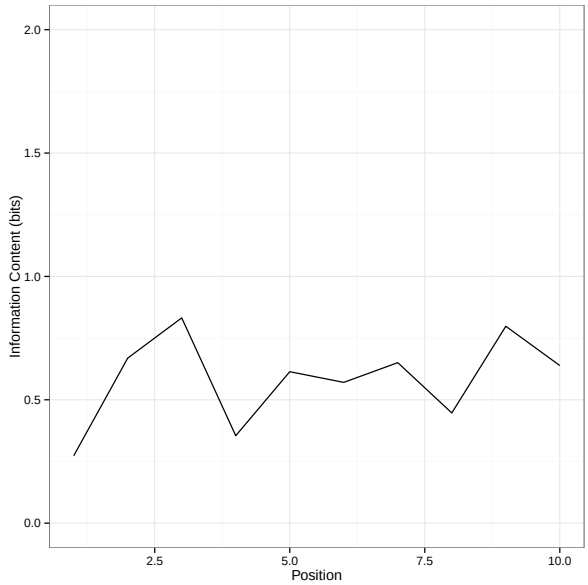
	1	2	3	4	5	6	7	8	9	10
A	−0.08	−2.58	1.47	0.50	0.95	0.69	1.22	0.22	1.22	−0.12
C	1.05	0.69	−2.44	−1.87	−4.01	−3.65	−2.42	−2.63	−4.42	−1.74
G	−2.22	−3.69	−1.69	−2.18	−3.56	−3.39	−3.96	−2.02	0.42	1.53
T	−0.24	1.04	−1.71	0.73	0.48	0.76	−0.13	0.98	−3.16	−2.11

(c) Specificity Matrix

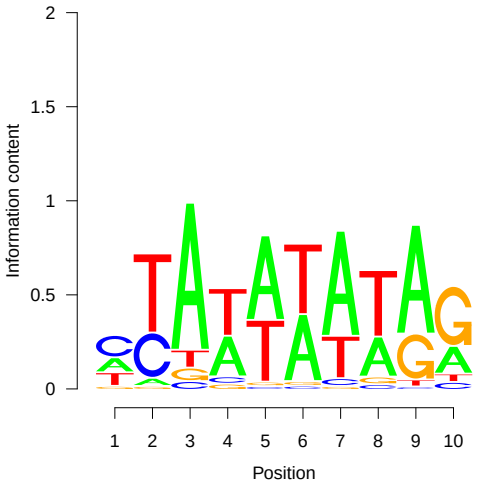
Table 3.2: **(a)** Number of occurrences of each base at each position of the 1875 aligned sequences (see Section 3.3). The column sums equal 1875. **(b)** The counts divided by the total sum. This is the fraction of each base at each position. **(c)** Logarithms (natural base) of those fractions divided by the background frequency. The minimum and maximum scores of the PWM are −31.311 and 10.950.

	1	2	3	4	5	6	7	8	9	10
A	75	22	706	65	158	72	384	138	705	265
C	599	157	81	42	20	19	44	54	5	37
G	49	12	35	10	10	14	26	14	233	503
T	277	809	178	883	812	895	546	794	57	195

Table 3.3: The *Count Matrix* as obtained from the software MEME. The input was the same set of high confidence binding site used to construct Table 3.2.



(a) The Receiver Operating Curve



(b) The Statistical Significance of thresholds

Figure 3.2: The information content plot (left) of the PWM in Table 3.2. The sequence logo plot (right) is a graphical visualization of the most conserved bases at each position

significantly to the aggregate score R and how likely is it to achieve score R from the background nucleotide composition. Second, typical application of a PWM model is to scan large number of sequences to identify novel binding sites. The sensitivity and specificity of the model is affected if high scoring sequence matches have a high probability of occurring by random chance. The probability of a PWM match occurring by chance depends on the target sequence as well as the background nucleotide composition [10]. Therefore, running statistical significance tests is imperative in assessing the performance of the model.

The statistical significance tests also entails a biological interpretation. Recall that the binding energy of any w-length sequence u is $\boldsymbol{W} \cdot \boldsymbol{u}$ which is a measure of how close the sequence u is to the consensus sequence (*motif*) determined by the PWM. For a PWM model to be viable, biologically significant sites must correspond to a high score [29, 54, 56, 10]. So we assume that there exists some threshold of binding energy (a *threshold score*) such that a sequence must have to have regulatory functionality[54, 56, 38, 29, 10]. We denote this threshold score α. So, given a threshold score α, we say that the PWM $\boldsymbol{W}$ has an occurrence in a target sequence S if there exists a w-length subsequence u such that $R = \text{Score}_W(u) \geq \alpha$. The question of interest is, then, how to choose the optimal threshold value α. A fair and valid assumption is that the optimal threshold value should be such that it minimizes the number of false positives while maximizing the

number of true positives.

A PWM model is applied through the score function (3.1) where the entries of the PWM are the logarithms of a likelihood ratio, or *log-odds*. Given a PWM $\boldsymbol{W}$, and a w length sequence u, we have two hypothesis [10, 56, 66]

- **Null θ_{yes}:** The w-length sequence u belongs to the model with position-specific constraints.
- **Alternative θ_{no}:** The w-length sequence u belongs to the background model with no position-specific constraints.

The likelihoods for observing sequence u under these hypothesis is given by

$$\pounds_{yes} = \Pr(u \mid \theta_{yes} = W(u_1, 1)) + W(u_2, 2)) + \ldots + W(u_w, w) \tag{3.13}$$

$$\pounds_{no} = \Pr(u \mid \theta_{no}) = \prod_{b \in \Sigma} \pi_{bg}(b) \tag{3.14}$$

where u_i is the nucleotide b at the i'th position of the sequence u and π_{bg} is the background probability. The logarithm of the ratio $\log_2(\pounds_{yes}/\pounds_{no})$ leads precisely to the score function (3.1). It is clear that the background frequencies as well as the pseudocounts play roles (somewhat critical roles [66]) in determining statistical significance.

In what follows, we consistently apply the PWM from Table 3.2 to provide examples from the theory. Denote this matrix as $\bar{\boldsymbol{W}}$. We use this notation to keep $\boldsymbol{W}$ as a dummy variable representing any PWM model.

Statistical Significance of Individual Positions The statistical significance of PWM's individual position scores can be assessed by χ^2-tests with the type-I error rate controlled using false discovery rates [66]. In our particular case, for $\bar{\boldsymbol{W}}$, Table 3.4 show 10 different χ^2 tests using $\pi_{bg}(b) = \{A = 0.292, C = 0.208, G = 0.208, T = 0.292\}$. Let the error rate be ν_0, then the rejection region is given by

$$\nu = 1 - \left[(1 - \nu_0)^{\frac{1}{N}}\right] \tag{3.15}$$

where $N = 1875$, the size of the aligned dataset used to construct the PWM. Setting $\nu_0 = 0.05$, we have that $\nu 2.74 \times 10^{-05}$. Even after applying p-value adjustment method (*Benferroni*), we reach the conclusion

	A	**C**	**T**	**G**	χ^2	Pr
1	518	809	84	464	704.572	0.0000000
2	91	631	30	1123	1466.790	0.0000000
3	1515	72	121	167	2418.963	0.0000000
4	773	107	86	909	773.886	0.0000000
5	1054	24	33	764	1224.450	0.0000000
6	880	31	37	927	1114.954	0.0000000
7	1276	73	25	501	1572.553	0.0000000
8	637	63	96	1079	1026.405	0.0000000
9	1272	18	524	61	1791.890	0.0000000
10	502	117	1130	126	1923.481	0.0000000

Table 3.4: The Count Matrix for the MEF2 transcription factor from Table 3.2. The χ^2 test is performed for each position against $\pi_{bg}(b) = \{A = 0.292, C = 0.208, G = 0.208, T = 0.292\}$. The p-values are all 0.

that that the frequency distribution of all sites deviate significantly from that of the background frequency distribution.

Computation of Score Distribution The statistical significance of our PWM is analyzed by its score distribution. The score assigned to a sequence (equation (3.1)) by $\boldsymbol{W}$ is a measure of the degree of similarity between the sequence and the PWM. Figure 3.3 shows, from $4^10 = 1048576$ 10-length long DNA sequences, the number of sequences that are below various binding energy values for $\bar{\boldsymbol{W}}$. As expected, the number of sequences equal or below a threshold value follows an exponential distribution. In other words, sequences with high affinity (low energy) follows the exponential distribution (Figure 3.3) [38]. The number of distinct scores of a position weight matrix of with non-negative integer entries and length w is bounded above by $\sum_{i=1}^{w} \max W(i,b) \colon b \in \Sigma$. In practice, however, matrices often are real-valued, such as $\bar{\boldsymbol{W}}$. For such a matrix, the theoretical maximum number of distinct possible scores is $|\Sigma|^w$. The histogram of all possible scores of $\bar{\boldsymbol{W}}$ is shown in Figure 3.1. As expected, the scores approximately follow a normal distribution [10, 59, 66] with mean $-10.514(0.00562)$ and standard deviation $5.76(0.003977)$. The cumulative distribution $C(R)$ represents the probability for an individual match to score $\leq R$. However, the cumulative distribution function $C(R)$ is not yet the proper one. Suppose we apply the PWM model on a target sequence S of some length. For example, scanning a sequence S of length 10 has a score $R = 10.95$. The 99'th percentile confidence limit, using the normal distribution with mean -10.514 and standard deviation 5.76, is 4.32 which implies that a score of $R = 10.95$ is statistically significant at 0.01 confidence level. However, consider

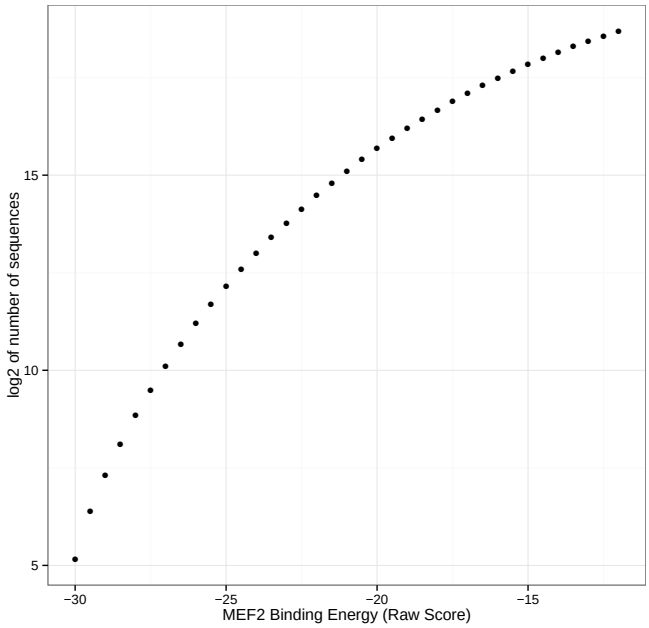
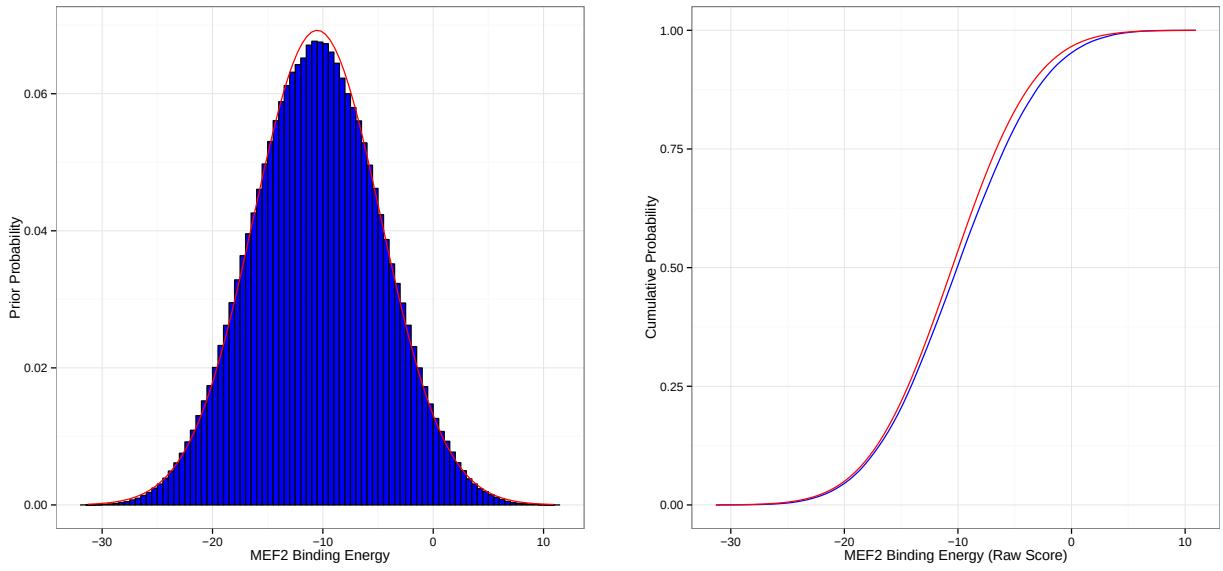


Figure 3.3: The $\log_2$ of the number of sequences (from all $4^10 = 1048576$ 10-length DNA sequences) that are equal or less than the binding energy calculated using PWM $\bar{\boldsymbol{W}}$, indicated on the x-axis.



(a) Probability Density Function

(b) Cumulative Distribution Function

Figure 3.4: (a) Prior distribution of binding energy for the MEF2 transcription factor PWM (Table 3.2). In addition, a fitted normal distribution (red) with mean $-10.514(0.00562)$ and standard deviation $5.76(0.003977)$ (b) The cumulative distribution function where the red curve is from the fitted distribution and the blue curve is the empirical distribution.

when the length of the target sequence is 1000 nucleotides. In this case, the PWM model returns 991 scores for the 991-length subsequences in S. To assess the statistical significance of these 991 subsequences (and their scores), we compute the cumulative probability and density functions of all matches. We can derive the cumulative probability and density functions by performing many sampling experiments [66] using the normal distribution density function Φ. Denote the scores of the N subsequences by $R_1, R_2, R_3, \ldots, R_N$. Let the maximum score be R_{max}. The probability of getting a score R less than R_{max} is given by

$$C(R \leq R_{max}) = \int_0^{R_{max}} \Phi(R)\, \mathrm{d}x$$

There are $N-1$ R_i values that are $\leq x_max$. We can define a density function F:

$$F(R_{max}) = N\, \Phi(R_{max})\, C(R \leq R_{max})^{N-1} \tag{3.16}$$

Figure 3.5 show the plots of F for increasing length target sequences using $\Phi(-10.514, 5.76^2)$. As expected, the expected best score distribution tends to the *Extreme Value Distribution (EVD)* [10, 66]. This is intuitive since R_{max} is the extreme value of N R_i values, so it is natural to see the EVD. Comparing the curves for for various sequence length values (N) and the fitted normal distribution, the distribution of F has been condensed significantly and shifted to peak at R_{max}. In general, as the width of the target sequence increases, the probability density resembles the extreme value distribution [10]:

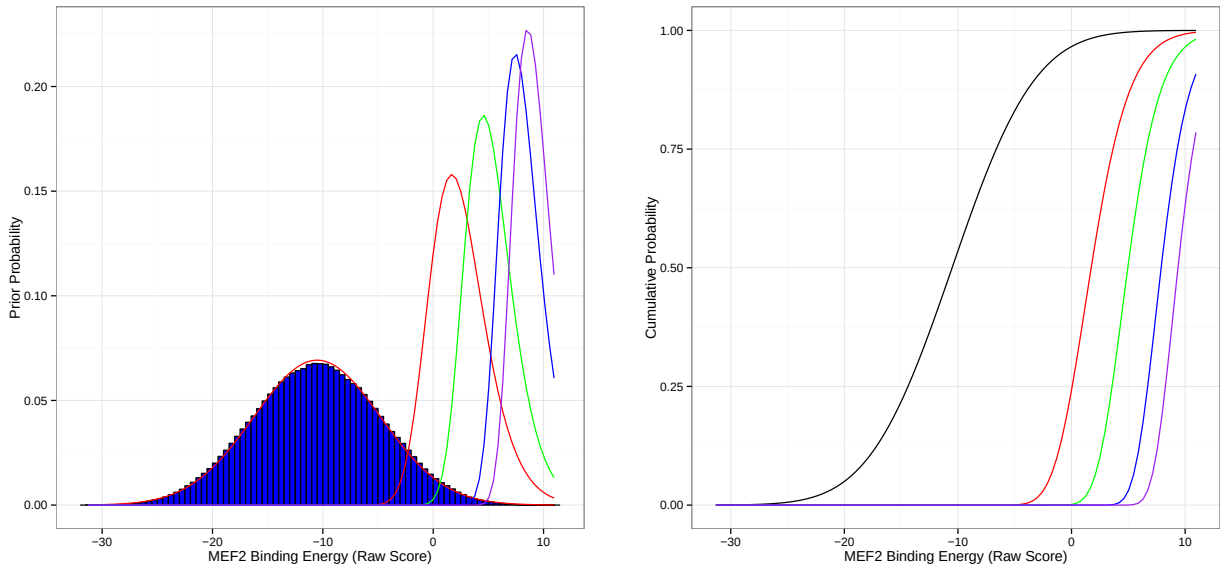
$$g(z) = \frac{1}{\beta} e^{\frac{z-\mu}{\beta}} \exp\left[-e^{\frac{-(z-\mu}{\beta}}\right]$$

The EVD is used to assign statistical significance to the sequence that score R_{max}. The probability of having one sequence score $\geq R_{max}$, ie the statistical significance, is give by the complement, that is

$$\Pr(R_{max} \geq R_{obs} = \int_{R_{obs}}^{\infty} F(R_{max})\, \mathrm{d}R_{max}$$

Consider an experiment in which we apply $\bar{\boldsymbol{W}}$ to a target sequence of 1000 nucleotides. The model returns 991 scores with the maximum score, say, $R_{max} = 8.32$. The probability of observing this $R_{max} \geq 8.32$ is 0.66, which is not significant at all.

The EVD is a common distribution in the realm of bioinformatics. The distribution is useful for predicting the chances of extreme outcomes. It is suggested that the extreme value distribution (also referred to as



(a) Probability Density Function

(b) Cumulative Distribution Function

Figure 3.5: (a) The red, green, blue, purple curves is the density function F of target sequence lengths N: $50, 200, 1000, 2000$, respectively. The curves were plotted using $\Phi(-10.514, 5.76^2$ (b) The cumulative probability functions for the extreme value distribution

the *Gumbel Distribution* may all govern ungapped pairwise alignment problems[10]. As such, it is used in the popular bioinformatics tool, *BLAST*, to attach statistical significance to an alignment score [10, 66]. In general, determining the distribution of PWM scores is necessary in order to assess the statistical significance of matches and also to estimate the expected rate of false positives.

Statistical Significance of the I_{seq} So far, we've detailed methods to assess the significance of the scores as well as the significance of the relative binding energy, given by the entries of a PWM model. However, we can assess the significance of the model itself. Recall that the information content (3.12) measures the relationship between the ΔG of a functional TF binding site to the ΔG of a TF binding to an arbitrary DNA sequence [56, 35, 54]. Statistically, it is a measure of the distance from the background frequency distribution [35]. Hertz and Stormo [35] suggests that $e^{-I_{seq}}$ is an upper limit to the expected frequency of the individual bases within an aligned dataset. From the definition (3.12), it is clear that this statistic depends on the pseudocounts as well as the background frequency. The statistic I_{seq} can have additional interpretations if the p-value is calculated. In this context, the p-value is the probability of observing an aligned dataset has an observed information content greater than I_{seq}. A theoretical study can be found in

Touzet and Varré [59] and Hertz and Stormo [35]

3.5 Results and Simulations

A PWM is a probabilistic model to discover regulatory regions and specific binding regions for proteins/DNA complexes. The PWM $\bar{\boldsymbol{W}}$ in Table 3.2 is specific for discovering binding sites for the MEF2 transcription factor. Like any other model, we are interested in the performance of our model and its ability to accurately discriminate between functional binding sites and non-specific binding sites. We perform sensitivity and specificity analysis and characterize the performance of the model by ROC analysis. When $\bar{\boldsymbol{W}}$ is applied to genomic regions of interest, returns DNA sequences (or *matches*) of length $w = 10$ and their associated score R. Let the set of all PWM matches be U. A sequence $u \in U$ is a *potential binding site* if its score, R, is greater than or equal to some threshold value α. It is important to note that the raw score R is not particularly informative and quite arbitrary. Therefore, we apply a simple transformation which maps R to a percentile score P given by

$$P = \frac{R - \min(\boldsymbol{W})}{\max(\boldsymbol{W}) - \min(\boldsymbol{W})}$$

where the *min* and *max* are the minimum and maximum scores of the PWM [33, 54, 29]. In particular, the minimum and maximum scores of $\hat{\boldsymbol{W}}$ are -31.311 and 10.950 respectively. In this context, a potential binding site is such that the percentile score

$$P > \alpha$$

where α is a user-defined threshold corresponding accordingly to the percentiles. Regardless of using R or its percentile score P, the statistical significance can be computed using Equation 3.16. More precisely, this gives us the P-value of the score: the probability is the random expectation of observing a raw score of R or greater [55]. The p-value can be estimated theoretically based on an extreme value distribution or empirically using by fitting a distribution on all possible scores $\hat{\boldsymbol{W}}$ can achieve [10].

3.5.1 Preliminary Accuracy

A preliminary performance analysis of $\hat{\boldsymbol{W}}$ was first performed by scanning a collection of 17 short target sequences. These target sequences have sufficient evidence of containing MEF2 binding sites and have been

studied in a number of organisms [29], which adds validity to the fact that the binding site for MEF2 is highly conserved amongst mammals [7, 62]. The accession numbers, description, and the start position for the binding site is described in Table 3.5. There is a natural expectation that, if $\hat{\boldsymbol{W}}$ has captured a high enough information content from the 1875-block alignment data (section 3.3), individual binding sites will be relatively *high* scoring. Indeed, there are only seven sites which scored less than $P = 0.90$ and only three sites that scored less than $P = 0.85$ (Table 3.5). Figure 3.6 describes the sensitivity analysis.

To assess the predictive power of $\hat{\boldsymbol{W}}$, neighborhoods of 400 nucleotides about each known binding site were scanned with $\hat{\boldsymbol{W}}$. The neighborhoods were selected so that the binding site is arbitrarily near the center. Using a sliding window of length $w = 10$, the percentile scores P of all subsequences in each neighborhood are calculated. All matches with a score $\geq$ a varying threshold score α are classified as positives and all matches scoring $\leq \alpha$ are classified as negatives. Since the true location of the binding site was known *a priori*, the performance of the model can be measured by ROC analysis (chapter 6). For increasing threshold values, the sensitivity (fraction of actual sites located) and specificity (1 - fraction of false identification) at each threshold value is computed and plotted. The corresponding ROC curve is plotted in Figure 3.7. The area under the curve is 0.9113 with 95% CI: $0.8535 - 0.9479$. The high AUC value suggests that the PWM $\hat{\boldsymbol{W}}$ is highly predictive of binding sites, relative to a small search space.

Receiver Operating Characteristics A Receiver Operating Characteristics (ROC) graph is a tool to visualize, organize, and evaluate classifying models based on their performance. ROC curves are two-dimensional graphs in which the true positive rate is plotted against the false positive rate. Many classifiers are designed to produce a decision (`YES` and `NO` , `TRUE` and `FALSE`) on each instance. Applying such a classifier to a dataset returns a single confusion matrix corresponding to a single ROC point in the ROC space. However, some classifiers such as a Position Weight Matrix return a score of each element in the dataset. Such a scoring classifier can be used with a threshold β to convert to a binary classifier. If the classifier score output is larger than β, the classifier returns `TRUE` for the instance, otherwise `FALSE` . Conceptually, we may imagine varying a threshold value from ∞ to ∞ and tracing a curve through the ROC space. A brief exposition is provided in chapter 6. Further analysis and efficient construction of ROC curves are well reviewed in Fawcett [27] and [47].

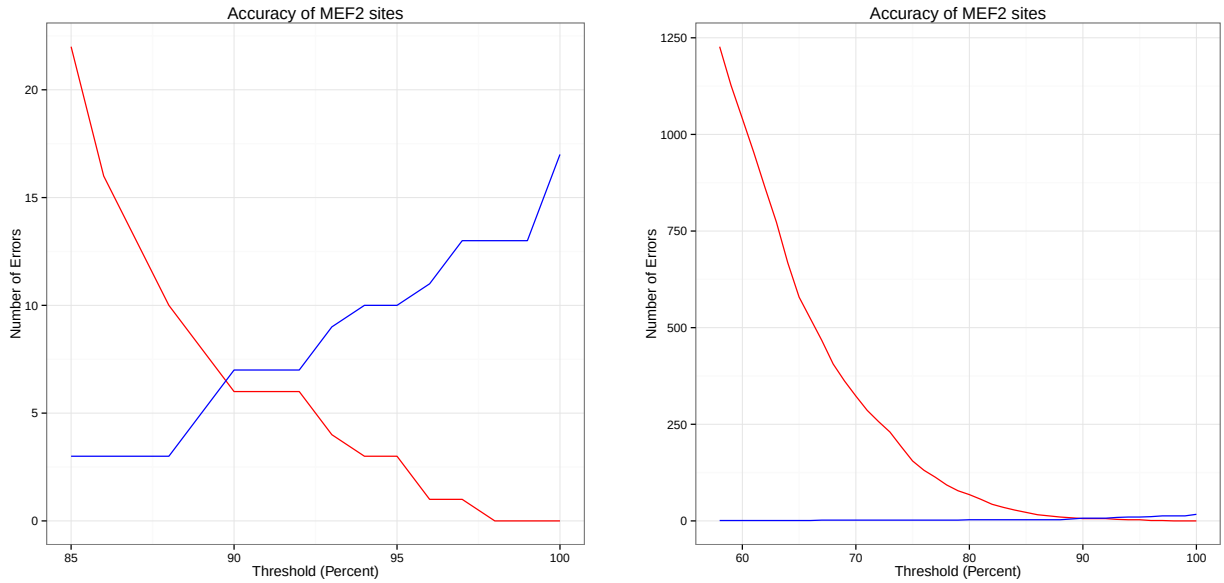
Accession	Description	Site Center	Site Start	Score
X06351	human aldolase A gene	1985	1981	0.994
X04260	R.norvegicus gene encoding aldolase A,	450	445	0.994
M62404	mouse cardiac myosin heavy chain gene	280	276	0.923
K01464	rat cardiac myosin heavy chain	280	276	0.791
M63391	human desmin	2286	2281	0.921
Z18892	mouse desmin	118	115	0.584
X58489	human GLUT4 enhancer	689	685	0.993
L36125	rat GLUT4 enhancer	1751	1747	0.993
M21487	human MCK enhancer	1772	1767	0.897
M27092	rat MCK enhancer	463	458	0.954
X14726	rat MLC 1/3	531	529	0.669
M37984	human slow/cardiac troponin C	2562	2557	0.962
J04971	mouse slow/cardiac troponin C	1904	1899	0.962
M80829	rat cTnt	912	908	0.893
X62155	human myogenin	1067	1063	0.882
M95800	mouse myogenin	1506	1503	0.882
M55673	human PGAM-M	1660	1657	0.939

Table 3.5: Natural sites taken from reference [29]. The table shows the center of the binding site and the score of the binding site using the PWM define in 3.2

3.5.2 Large scale analysis of the PWM model

Figure 3.6 characterizes the behavior of the PWM (Table 3.2) on relatively short sequences. In fact, the model exhibits high predictive power characterized by an AUC of 0.9113. Practically, it is of interest to determine the predictive power of the PWM model when the search space is large and no information is known about the true binding site location. In other words, given a sequence or arbitrary length for which no information exists on the true location of the binding site, a potential binding site is such that its score is above some threshold value α. Typically, optimal threshold values are such that sensitivity and specificity are maximized, in the context of the given sequence. Other criterion, such as cost-benefit analysis, can also be utilized for identifying an optimal value. Most often, different criterion return different thresholds and thus the choice of choosing an *optimal* one is quite arbitrary. In this section, we consider methods to find optimal threshold values and analyze the predictive power of the PWM $\hat{\boldsymbol{W}}$ on large sequences. All analysis is performed on the mouse genome (build 37 assembly by NCBI).

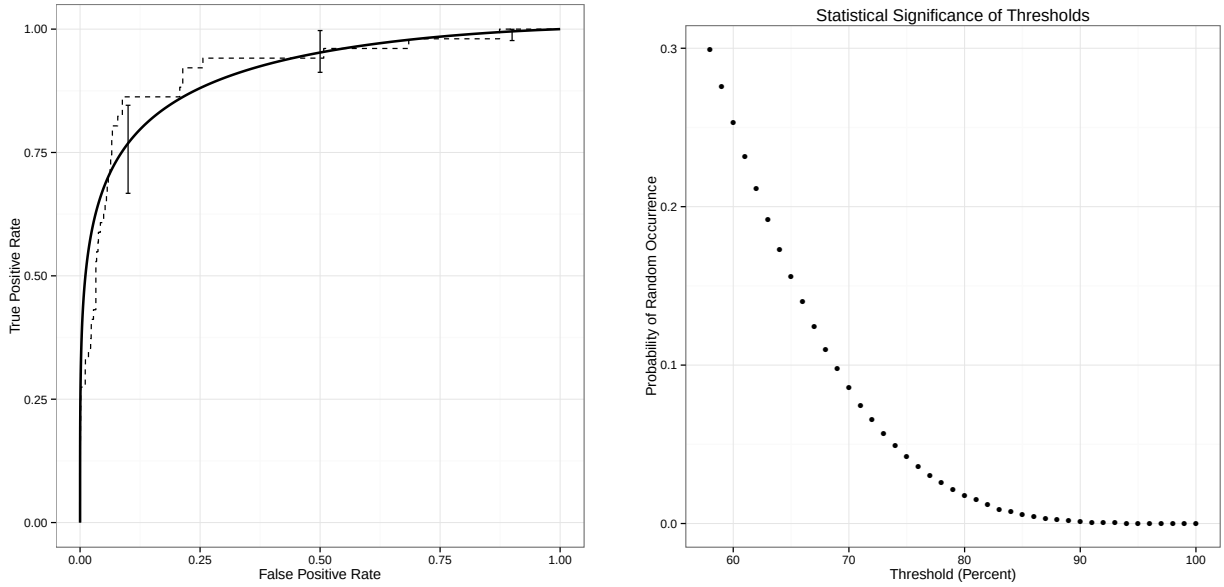
In general, for PWM models, as the specificity increases the sensitivity decreases thus making the decision of choosing an optimal score threshold quite arbitrary. Returning previously to when PWM is applied on 400nt



(a) Threshold over 90%

(b) Threshold over 50%

Figure 3.6: The red curve represents the number of false positives. The blue curve represents the error rate, that is the true sites missed by the model. A search requiring a perfect match will result in no false positives but also miss all the true sites. The optimal threshold value is the value for which we minimize the number of errors and false positives.



(a) The Receiver Operating Curve

(b) The Statistical Significance of thresholds

Figure 3.7: The ROC curve for scanning 400nt sequences with the PWM; AUC value of 0.6575 with 95% CI: $0.627 - 0.6897$. Each 400nt sequence contains only one true site.

neighborhoods of Table 3.5, consider when all matches with score $P \leq 0.90$ are discarded. For the remaining 16 matches, 4 additional sites were found while missing 5 true ones. If the four additional sites are spurious, the false-positive prediction rate is one site per 1600 bases. In contrast, at a threshold of 99%, we have zero false positives (perfect specificity) but 13 true sites are missed thus resulting in poor sensitivity. Several methods have been developed for selecting threshold values for classifier models. In fact, the `R` package `OptimalCutpoints` implements over 20 different criterion for selection optimal cut-off points. Indeed, this demonstrates that picking an optimal cut-off point is arbitrary. Nonetheless, in the absence of experimentally verified binding sites, it is imperative to establish such a cut-off point. To this end, considering the criterion based on simultaneously maximizing sensitivity and specificity returns a percentile threshold of $\alpha = 0.65$ (raw score: -3.8111), with a sensitivity of 0.823529 and specificity of 0.916161. Recall that scores from random sequences follow an approximate normal distribution, in particular with mean -10.514191 and standard deviation 5.760025 for $\hat{\boldsymbol{W}}$. The upper 99% confidence limit is 4.322586 which implies that the score of -3.8111 is not statistically significant. Similarly, if using the criterion that the optimal threshold is the ROC point that is closest to the point $(0, 1)$ also returns $\alpha = 0.65$. The criterion based on the cost-benefit methodology by calculating the slope of the ROC curve at the optimal threshold value given by

$$S = \frac{1-p}{p} CR = \frac{1-p}{p} \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}}$$

where $C_{(\cdot)}$ are the costs of false positive, true negative, false negative, and true positive decisions and p is the *disease prevalence*. This method returns the value $\alpha = 0.96$ (raw score: 9.260), with sensitivity 0.1176 and specificity 0.9998. This score is statistically significant at the 0.01 level. In fact, the upper 99.9% limit is 8.439 which implies that α is significant at the 0.001 level.

Selecting an optimal threshold offers the ability to analyze the genome-wide predictive power of $\hat{\boldsymbol{W}}$, ie discover novel binding sites. The sheer size of the genome combined with the fact that binding sites are degenerate proves to be a difficult task. Consider a 10 million nucleotide sequence with a background model of equal nucleotide frequency $\pi_{bg}(b) = 0.25 \, \forall b$. The probability of a length 10 sequence being a perfect match to our consensus sequence is $0.25^{10} = 9.536743 \times 10^{-7}$. Although a small probability, scanning a 10 million nucleotide sequence results in 9.25 perfect matches entirely by chance. Similarly, scanning the entire mouse genome, which contains about 3×10^9 nucleotides, results in 2700 perfect matches entirely by chance. It is clear that biologically significant PWM matches will be overwhelmed by the chance matches. Therefore, applying the PWM to the entire genome without removing a considerable amount of the sequence will result

in a high false positive rate and overall poor performance of the model. The criterion used to remove sections of the genome from consideration is based upon a biological gold-standard. This gold-standard is created by using high-confidence experimental data for the MEF2 binding sites obtained from Wales et al. [62]. Their study identified only 2797 binding peaks (2783 after mm10 $\rightarrow$ mm9 conversion), which is low compared to some other studies done and relative to the number of genes. This could be because the experiment's methodology used ChIP-exo rather than ChIP-seq [62]. There may also be a biological reason: MEF2 has four isoforms in mammals but the study [62] only considered the MEF2A isoform. Furthermore, the time point for cell differentiation was set to 48 hours. Perhaps MEF2A has different or more binding targets in fully differentiated cells. We setup a biological gold standard by following the methods of Cuellar-Partida et al. [12]. This gold standard begins with the experimental data of the transcription factor of interest, retrieved from Wales et al. [62]. This gold standard first removes from consideration all genomic positions that are deemed to have low evidence of potential sites. The potential sites in the remaining genomic positions are labeled positive or negative based on strong ChIP-seq evidence for or against occupancy by the transcription factor. Potential sites are marked positive if they fall within a ChIP-seq peak. Potential sites are marked negative if they are not within a peak.

Genomic regions of interest Generally, most transcription factors bind to either the enhancer or the promoter regions of the genes they regulate [25]. Depending on the tissue, cell line, and the transcription factor, the transcription of genes is either up-regulated or down-regulated. In particular, MEF2 usually binds to the promoter regions of the genes they regulate [7]. Therefore, it is reasonable to apply the PWM model to promoter/enhancer regions only. Currently, there does not exist a canonical size for regulatory regions, such as promoters of genes. Indeed, regulatory regions can be located over one million nucleotides away from their target gene. In addition, there may be other biological factors in between a regulatory region and its target gene[7]. Considering these challenges, for all genes in the mouse genome (mm9), we define regulatory regions of increasing lengths: $10,000$, $20,000$, $50,000$, and $100,000$. In particular, these are regions upstream (and 5000nt downstream) of the transcription start site of the canonical isoform of the target gene. Here, the canonical isoform is taken from the known table of the UCSC known genes track. Increasing length regions also allows for analyzing the predictive power of the PWM at a genome-wide level, after all the ultimate objective is to apply such models to discover novel binding sites.

The NCBI build 37 mouse genome contains $21,677$ genes. Scanning 100knt regions of all $21,677$ genes results

in an infeasible sequence size of $2,167,700,000$. Therefore, in further effort to reduce the search space for the model, note that MEF2 is a transcription factor only specific to muscle related genes. Thus scanning regulatory regions of all genes is not necessary, and a considerable amount of genome not associated with MEF2 can be removed. In addition to determining binding site locations, Wales et al. [62] also associates each binding site with its target gene. To this end, there are 3121 genes associated with the MEF2 transcription factor. Furthermore, 2730 (out of 2797) of the peaks lie within a 100nt regulatory region of all 3121 genes.

Applying the model on varying length regulatory regions Recall that a PWM model with length w works by scoring each sliding window of length w. For example, a 10-length PWM applied on a 200nt sequence returns 191 total scores. Therefore, 3121 genes with regulatory regions of size $100,000$ implies the search space for the PWM model are $312,100,000$ subsequences. In a further effort to reduce the search space, a smoothing process was implemented. For each gene, the defined genomic region was split into 200nt intervals where the score of each interval was set to the maximum score of all the subsequences within the interval. The choice of a 200nt interval length was decided because the average length of a binding site peak from [62] was 194nt, and furthermore, to keep consistency of the nucleotide resolution of bins in chapter 2.

First, $\hat{\boldsymbol{W}}$ was applied to 10knt regulatory regions. From a total of $214,526$ intervals, 560 were *positive*, as per the gold standard. The lowest interval score was 0.5613 and a mean score of 0.8782. Applying ROC analysis returns AUC of 0.5514 (95% CI: $0.5248 - 0.5779$). Increasing the length of regulatory regions to 20knt returns $359,476$ intervals with 806 positive intervals. The lowest interval score was 0.5295 and a mean score of 0.8753. Applying ROC analysis returns AUC of 0.5518 (95% CI: $0.5299 - 0.5736$). For regulatory region lengths 50knt and 100knt, the results are similar. The lowest scores for intervals are 0.4939 and 0.4939 with mean values of 0.8765 and 0.8770, respectively. The associated AUC values are 0.5566 (95% CI: $0.5395 - 0.5738$) and 0.5629 (95% CI: $0.5486 - 0.5772$), respectively. It is clear that the large number of intervals with a low number of verified positives results in poor accuracy of the model. The mean score of ≈ 0.87 is not significant at the 0.001 level under the normal distribution, regardless of regulatory region size. ROC curves are plotted in Figure 3.8.

The low predictive power of the PWM model can be mitigated by considering the biological significance of scores. Recall that PWM scores are estimates of the free energy of the transcription factor binding to sequence sites. Therefore, it is reasonable to expect that the predictive power of the model can be vastly improved if all PWM matches below a particular threshold value α are discarded. In other words, potential binding

sites are such that their score $\geq \alpha$. Although, the choice of an optimal α is quite arbitrary, nonetheless, we set threshold value $\alpha = 0.96$ corresponding to cost-benefit criterion described above. Indeed, the PWM model exhibits improved accuracy when low-scoring matches are discarded. In particular, for 10knt regions, discarding matches below the threshold yields $7,119$ intervals, out of which 73 are identified positive. Most notably though, there is an increase in the AUC from 0.5514 to 0.6879 (95% CI: $0.6144 - 0.7615$). Similarly, for 20knt regions, the AUC improves to 0.6973 (95% CI: $0.6335 - 0.7613$). For 50knt and 100knt regions, the new AUC values are 0.6958 (95% CI: $0.6415 - 0.75$) and 0.6671 (95% CI: $0.6209 - 0.7133$).

3.6 Discussion

Identifying binding sites of transcription factors is a key problem in bioinformatics, and elucidates cellular gene regulatory mechanics. The consensus sequence model is the simplest model for predicting binding sites. However, since binding sites are degenerate in nature, consensus sequence based models are often an unsuitable approach. A Position Weight Matrix (PWM) is a probabilistic model that expands on the principles of the consensus sequence. The PWM model assigns a score to a DNA string of length w where w is the width of the matrix. The score is viewed as a log likelihood ratio for the hypothesis that the site will be found under the frequency model $f(b_i)$ versus the background model $\pi(b)$ [29]. High scoring matches are labeled significant (ie, potential binding sites) if their score is above a user defined threshold α. The score of a site lends to a biological interpretation as well. The elements of the PWM model are interpreted as estimates of the free energy for the protein binding to the site. Under the additivity assumption, the total binding energy of a site is, thus, the sum of its individual scores for each base at each position. It is expected that an acceptable PWM is such that biologically significant binding sites correspond to *high scoring* matches at the positions of the binding sites. However, there currently exists no consensus method for determining an optimal threshold *alpha*, such that sites with scores $\geq \alpha$ are deemed significant. Some studies calculate an approximate optimal threshold by using a test set of positive verses negative sequences, but this is difficult to do as obtaining negative sequences is not feasible. Therefore, statistical methods are required.

The necessity of using statistical methods was recognized by Claverie and Audic [10] and Xia [66]. There is sufficient literature of developing methods for assessing the statistical significance of scores. This is crucial as statistical significance presents a way to calculate performance of a model in the absence of a gold standard.

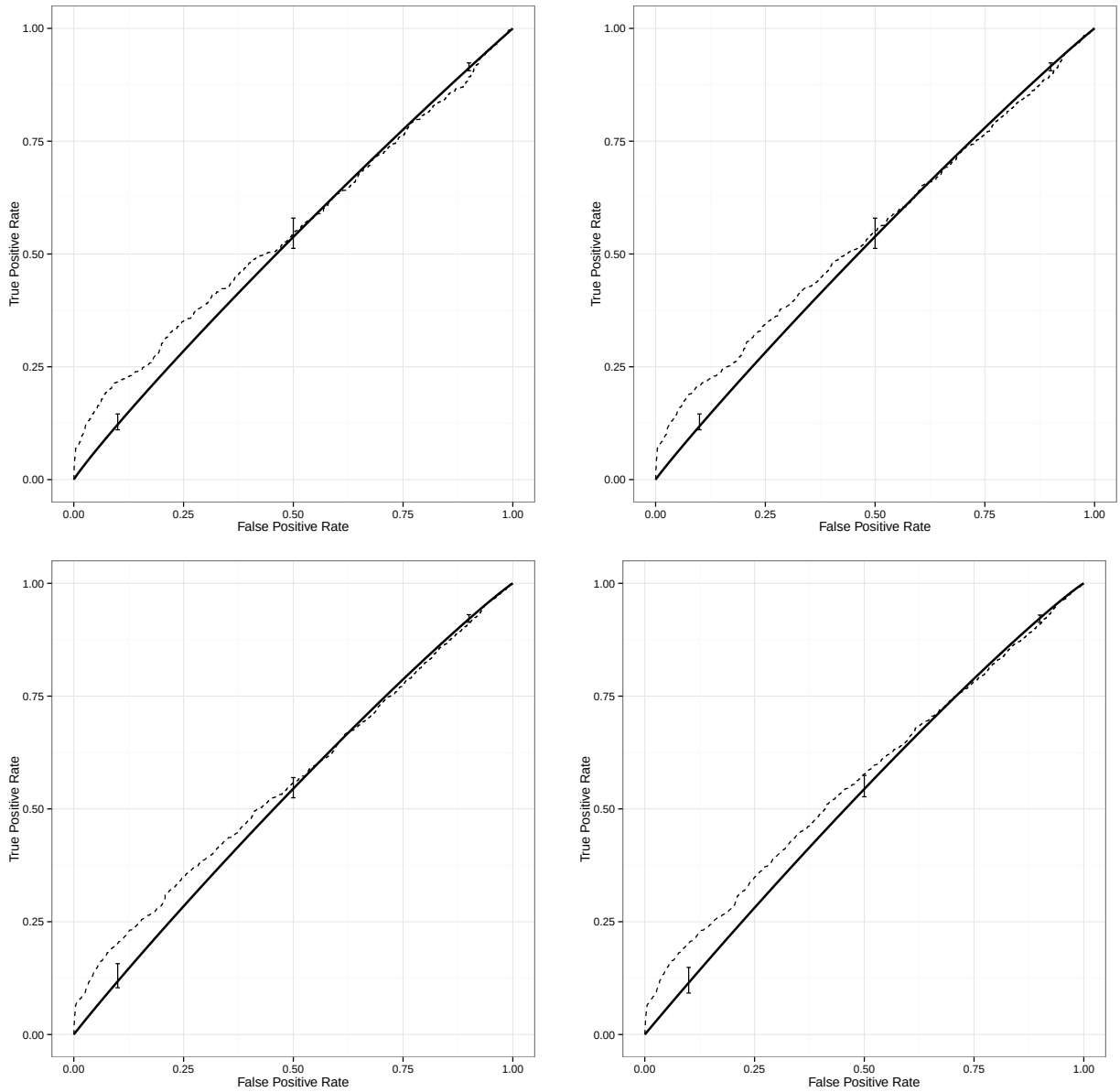


Figure 3.8: Empirical and smoothed ROC curves of the PWM model applied on genomic regions of length 10, 20, 50, and 100 kilo-nucleotides. The low AUC values of 0.5514, 0.5518, 0.5566, and 0.5566, respectively, suggests a poor accuracy of the PWM model.

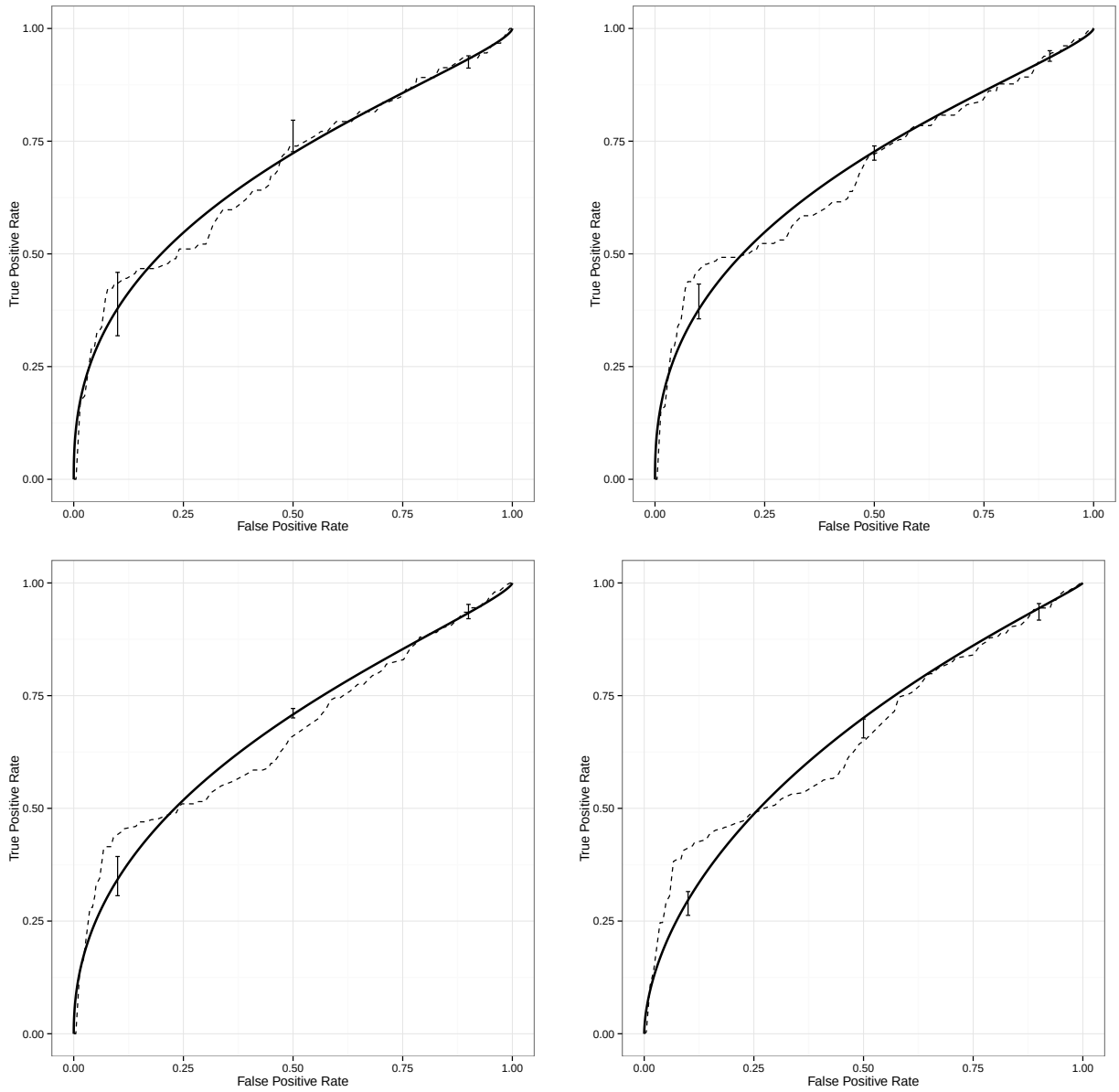


Figure 3.8: Empirical and smoothed ROC curves of the PWM model applied on genomic regions of length 10, 20, 50, and 100 kilo-nucleotides. All statistically insignificant ($p = 0.001$) matches were discarded. In particular, all PWM matches $\leq \alpha = 0.96$ were removed from ROC analysis. As expected, the performance of the PWM model improved considerably characterized by AUC values of 0.6879, 0.6973, 0.6958, and 0.6671, respectively.

Determining the statistical significance is inherently challenging due to the nature of biological processes. In particular, the repetitive nature of DNA can generate high scoring results in a non-functional site of the genome; conversely, low scoring results may be fundamental components in gene regulatory mechanics. Nonetheless, it is established that individual PWM results follow an approximate Normal distribution and extends to the Extreme Value distribution when the search sequence is increased. Despite the rigorous statistical methods and an appropriate biological interpretation, the PWM model is prone to high number of false positives and poor accuracy. The is exacerbated in mammalian genome since cis-regulatory elements such as binding sites can be kilobases away from their target genes, thus making it necessary to search large regions [57].

In addition to the theoretical analysis, the discriminatory ability of the MEF2 specific PWM model is assessed by ROC analysis. The performance of a PWM model can be assessed by its ability to classify true binding sites while minimizing the false positives. Ultimately, ROC analysis revealed poor performance of the MEF2 specific PWM model, characterized by low AUC values. We hypothesize that, although an attractive model due to its simplicity, various assumptions such as independence between nucleotides and independence between binding sites generates results that are not biologically significant, and thus a high false positive rate.

3.A Pseudocounts

The elements of a PWM are log-likelihood ratios of a base appearing at a certain position derived from a collection of high confidence, experimentally verified binding sequences. If this collection contains a small number of sequences, the count of nucleotides in each position may be skewed, and thus sway our belief from the neutral hypothesis. In the worst case, if the sample data is small enough, a nucleotide may not be observed at all for a particular position. This leads to zeros in the frequency matrix, and consequently infinities in the PWM when applying log function. In addition to the mathematical difficulty in dealing with infinities, there also exists biological motivation to remove them. In particular, assigning a probability of zero for an unobserved nucleotide b at a position i imposes too harsh of a penalty, but the variability of binding sites make it impossible to assign such a harsh penalty. To avoid a count of zero due to a small sample, common practice is to add *pseudocounts* to the observed counts. In essence this is a smoothing process. There currently exists no consensus on optimal pseudocount values, however Claverie and Audic

[10] and Nishida, Frith, and Nakai [46] provide methodologies to determine parameters that work well in most situations. We describe the two most common methods: *Constant Mode* and *Proportion Mode*.

Methods for computing pseudocounts The elements of a PWM are given as

$$M(b,i) = \ln\left(\frac{f(b,i)}{\pi_{bg}(b)}\right) \qquad i = 1 \to w \tag{3.17}$$

where $f(b,i)$ is the observed frequency of nucleotide b at position i in a block alignment of N sequences of width w. If a nucleotide is not observed at least once at any position i, there is an obvious problem with the above equation such that $\ln(0) = \infty$. Therefore, we modify the equation:

$$M(b,i) = \ln\left(\frac{F(b,i)}{\pi_{bg}(b)}\right) \qquad i = 1 \to w \tag{3.18}$$

where $F(b,i)$ is a **modified frequency**. We require $F(b,i)$ to follow properties:

$$F(b,i) \to f(b,i) = \frac{N(b,i)}{N}, \text{ and} \tag{3.19}$$

$$F(b,i) > 0 \text{ for } N(b,i) = 0 \tag{3.20}$$

This ensures that the modified frequency count $F(b,i) > 0$ for all positions i. We further impose two more restrictions on the modified frequency to ensure biological significance. An unobserved nucleotide b at position i should not correspond to a positive weight, hence if $N(b,i) = 0$,

$$M(b,i) \leq 0 \implies F(b,i) \leq \pi_{bg}(b)$$

where $\pi_{bg}(b)$ is the background frequency of nucleotide b. Conversely, if $N(b,i) > N\pi_{bg}(b)$, it corresponds to a non-negative weight, hence

$$M(b,i) > 0 \implies F(b,i) \geq \pi_{bg}(b)$$

Based on the required properties, a suitable formula for $F(b,i)$ is

$$F(b,i) = \frac{f(b,i) + \epsilon_b}{N + \epsilon} \tag{3.21}$$

where ϵ_b is the pseudocount for nucleotide b and

$$\epsilon = \sum_b \epsilon_b$$

We discuss two main ways to compute ϵ

Constant Mode In this method, the ϵ_b are identical for each nucleotide b. This method works well if the nucleotides in the target sequences for the PWM have a uniform background distribution. Equation 3.21 becomes

$$F(b,i) = \frac{N(b,i) + \frac{\epsilon}{4}}{N + \epsilon}$$

It is easy to check, given $\pi_{bg}(b) \approx 0.25$, that the properties above are satisfied. If the $N(b,i) = 0$ at a position i, the PWM elements are given by

$$M(b,i) = \ln \frac{\epsilon}{4(N+\epsilon)\pi_{bg}(b)}$$

Proportional Mode Typically, eukaryotic organisms tend not to have uniformly distributed nucleotides. In this case, the individual pseudocounts ϵ_b are functions of the a priori background distribution $\pi_{bg}(b)$. Equation 3.21 then becomes:

$$F(b,i) = \frac{N(b,i) + \epsilon \cdot \pi_{bg}(b)}{N + \epsilon}$$

It is easy to verify that this equation satisfies the properties required for F. If $N(b,i) = 0$ at position i, the element of the matrix becomes

$$M(b,i) = \ln \frac{\epsilon}{N + \epsilon}$$

3.B Literature Review

Transcription factors play a crucial role in gene regulation, by binding to specific DNA sequences in close proximity to their target gene. Predictive models constructed using a collection of known binding sites characterize the transcription factor's affinity to potential binding regions. In fact, Stormo [54] use a protein's specificity along with several known binding sites for the protein to develop a model for the specificity of the protein. Modeling a protein's specificity from example binding sites has been extensively studied over the

last decade [38]. The matrix based model, known as a Position Weight Matrix, provides a probabilistic as well as a realistic representation of protein/DNA interactions. In many cases, simple mononucleotide-based PWMS are adequate representations, but more complex matrices are easy to construct and can provide more information. A complete understanding of the theoretical, information based, construction of Position Weight Matrices in "Consensus patterns in DNA." (Stormo [54]) and "Neural networks for determining protein specificity and multiple alignment of binding sites." (Heumann, Lapedes, and Stormo [36]). Furthermore, a study by [22] presents an review of other information-theory based methods for constructing PWM models. Berg and von Hippel introduced a formal approach in modeling protein/DNA interactions by pure statistical mechanics [22, 35]. Staden [52]'s "Methods for calculating the probabilities of finding patterns in sequences." describes the use of probability-generating functions as a tool for pattern searching in sequences.

Besides the biological significance of PWM matches, the statistical significance of PWM matches is introduced by Claverie and Audic [10]. In the absence of experimentally verified data, computing the statistical significance offers a method in estimating the false positive rate and overall determine the accuracy of the model. More importantly, however, is that statistical frameworks offers an insight into choosing an appropriate score threshold for PWM matches. Claverie and Audic [10] and Xia [66] apply the framework of probability generating functions to PWMs and introduces both statistical theory as well as the numerical computation of the distribution governing PWM matches. In particular, the expected distribution of PWM scores tends to the Gumbel distribution, also known as the *extreme value distribution*

$$G(Z) = \frac{1}{\beta} \exp\left(-(Z-\alpha)/\beta\right) \exp -e^{-(Z-\alpha)/\beta}$$

Furthermore, a rigorous study by [59] provides methods for calculating P-values of PWM scores. The P-value is the probability that the background genomic frequencies can achieve a score larger than the score threshold α. The theoretical complexity proves that finding P-values is a NP-hard problem. The information content, described by [54] [35] [55] as a log-likelihood scoring scheme is a key statistic in calculating statistical significance of PWM matches. Hertz and Stormo [35] and Xia [66] review this statistic and provide numerical methods for estimating the P-value of an individual match's information content. They employ large-deviation statistics and provide an efficient algorithm for determining the moment-generating function to estimate the P-value as described by Staden [52]. In conclusion, these studies Erill and O'Neill [22], Stormo [55, 54], Claverie and Audic [10], and Hertz and Stormo [35] conclude that the information content is a sufficient measure of searching and quantifying the binding affinity of protein/DNA interactions.

Typically, sequence based models such as PWMs tend to have poor sensitivity and specificity. Techniques to improve the sensitivity and specificity of a PWM have been offered. [30] uses a modified algorithm, originally the Staden-Bucher algorithm, to increase a PWM's accuracy. Their modified algorithm uses a database of putative transcription start sites and returns a new 4-row (mononucleotide) and a 16-row (dinucleotide) PWM models. Their results show an improved PWM by suggesting optimal cutoff scores, but not necessarily the best PWM. A study by [42] implements a genetic algorithm to optimize a PWM. Their methodology maximizes the area under the ROC curves by incorporating prior information such as base conservation and other nucleotide information.

Word based algorithms, Machine learning techniques, based on genetic algorithms, and algorithms based on phylogenetic footprints are alternative toolkits in constructing predictive models of binding sites [13]. The study by Das and Dai [13], "A survey of DNA motif finding algorithms." takes on the difficult challenge to evaluate the performance of different motif finding algorithms. The difficulty in performance assessment arises from several sources. Mainly, this is because we do not have a clear understanding of regulatory networks and mechanisms, and thus it is difficult to obtain an absolute standard against which to measure performance of different algorithms.

4 Improving PWM model accuracy by integrating epigenetic modifications through a Hidden Markov Model

4.1 Introduction

It is well understood that transcription factors are key components in the spatio-temporal regulation of gene expression in mammals [9]. Identifying binding sites for transcription factors is a key step in modeling and elucidating regulatory networks. The current gold-standard in determining genome-wide binding sites of transcription factors is through experimental techniques [9]. Techniques based on chromatin immunoprecipitation (ChIP) followed by high throughput parallel sequencing (ChIP-Seq) or microarray hybridization (ChIP-chip) are wet-lab approaches to determine binding sites experimentally. These methods yield high-confidence binding sites, but only provide information on specific tissue types and conditions used in the experiment. Also, the vast majority of transcription factors have not been profiled genome-wide, thus finding their binding sites experimentally is infeasible [25]. Two main reasons that contribute to the lack of profiling is the cost of the experiment and the availability of the antibody for the transcription factor [38, 25]. Therefore, computational and mathematical models are necessary and inevitable. Computational approaches to predicting binding sites are based on pattern finding algorithms in computer science [25, 36]. The standard information-theory based Position Weight Matrix models are traditionally used to scan and locate binding sites. However, their use is limited due to a lack of reliable methods to assess statistical significance of PWM matches. In general, a PWM model returns many potential sites in which only a fraction are involved in gene regulation. In addition, the large search space, various assumptions of independence and nucleotide dependency results in a high number of false positives, and thus poor accuracy overall. As a result, the prediction of binding sites based on sequence data alone does not capture fully the spatio-temporal relationships in the protein/DNA interaction and thus the PWM model often exhibits poor accuracy.

In vivo, it is well documented that transcription factor binding mechanisms include more than sequence specific information [12]. These mechanisms rely on a multitude of biological factors to the limit the binding of transcription factors. Notably, local chromatin structure - the coiling of DNA, availability of secondary proteins, and epigenetics - plays a significant role in regulatory networks. Several post-translational covalent modifications of histones such as methylation, acetylation, phosphorylation, etc affect transcription factor binding in a complex and not-yet understood manner. However, numerous studies have shown empirically that several histone modifications are key components in gene regulation [24, 41, 3, 4, 9, 37]. Mounting evidence suggests that multiple histone modifications occur simultaneously so that certain recurrent and spatial combinations are directly associated with functional elements such as promoter and enhancer regions as well as cell-specific gene expression programs [3, 24]. It is expected the inferred locations of the binding sites by PWM models can be combined with other biological data such as gene expression to further gain insight into gene regulation and its dynamics. Many studies have developed methodologies that integrate multiple biological factors such as sequence data, evolutionary conservation, DNA clustering, gene expression levels, and functional similarity amongst TFs whose sites occur within close proximity [33].

In the preceding chapter, we constructed a Position Weight Matrix to identify binding sites for the transcription factor (TF) Myocyte-specific Enhancer Factor 2A (MEF2). This transcription factor is an activator, found in numerous muscle-specific genes, with a consensus sequence of $5' - YTA[AT]_4TAR - 3'$ [62]. The model was developed based on sequence data alone without additional information of chromatin structure. We found that the performance of the PWM model, characterized by the area under receiver operating characteristic curves, was poor. Although the area was ≥ 0.5, implying that our model is better than a completely random model, it is clear that the PWM requires improvement. Given successful applications of HMMs to capture chromatin information, we attempt to integrate PWM models and HM models to capture the relationship between epigenetic modifications and transcription factor binding sites. In this chapter, we combine the Hidden Markov Model learned on chromatin mark data in chapter 2 and the MEF2 specific PWM in chapter 3 in an effort to build a model for identifying novel transcription factor binding sites. The expectation is that a predictive model built from chromatin structure data as well as sequence data performs better then sequence-only PWM models.

4.2 Model Specification

The biological mechanisms underlying gene regulation are often at the level of transcription, such as the availability of RNA Polymerase, transcription factors, and other protein/DNA binding complexes. In particular transcription factors interact directly with the target gene's transcription complex, often in the promoter or enhancer regions of the target gene. However, increasing evidence shows that, in particular, histone modifications have been linked to gene regulation [67, 24]. The combinatorial interactions of histone modifications have given rise to the so called *histone code* hypothesis which suggests that the epigenome as a whole is a major mechanism in gene regulation networks. These modifications modulate the chromatin structure for the recruitment of TFs, enzymes or other proteins [9]. In general, however, the relationship between HM modifications and TF recruitment is unexplored. Recent statistical models have been developed that integrate the two biological mechanisms together to elucidate this relationship. Previous studies [9, 25, 57, 12, 44] have confirmed that both TF and HMs play a crucial role in predicting gene expression. These studies have developed models in which the information of HM have come from raw gene expression data. Hence, the accuracy of these models rely on experimental wet-lab datasets which are highly tissue and cell condition specific. Nonetheless, some of these models have suggested that TFs and HMs are more accurate predictors for transcription factor binding sites than using PWM models alone [9]. The model by Talebzadeh and Zare-Mirakabad [57] uses spatial positioning of nucleosomes harboring different combinations of histone modifications as an additional information source. Their results show that seven (our of 21) particular histone modifications have significant effect on transcription factor binding site predictions: H3K4me1, HeK4me2, H3K4me3, H4K20me1, H2BK5me1, H3K9Me1, and H3K27m1. Similarly, the model by Cuellar-Partida et al. [12] develops a novel heuristic method to integrate a prior distribution from epigenetic data. Their results also suggest that that histone modifications H3K4me1, H3K4me3, H3K9Ac, and DNase1 sensitivity conclusively improve TFBS prediction over a PWM model. Notably, their algorithm is now embedded in the popular bioinformatics *MEME Suite* toolkit. A large scale, genome wide study by Ernst et al. [25] incorporates 29 additional information sources, including distance from the nearest TSS, levels of histone modification, CpG islands, and evolutionary traits. In particular, they used 20 different histone modifications and concluded that combining these information sources improves prediction of TF bound regions.

Constructing an integrative model Our method works as follows: First, a genomic segment is assigned a score, constructed solely from epigenetic data, interpreted as the probability of *any* transcription factor binding to this genomic segment. In particular, this score is derived based on two main information sources: raw *binarized histone modification data*, and *HMM state data*. Specifically, this score is the output of logistic regression classifier (LRCs), trained on either information sources. It is important to know that this score is not specific to any *particular* transcription factor. In fact, Ernst et al. [25] denotes it as the *General Propensity Score*. It is suggested that the GPS is already highly predictive of true binding site locations, even when no sequence data is used [25, 9, 57]. Nonetheless, we integrate GPS with the PWM model in an effort to establish an improved TF predictive model. See Methods.

4.3 Results and Simulation

Modeling single histone modifications It is natural to first assess whether single histone modifications provide sufficient predictive power. Mapped tags for each of the nine histone modifications of the Asp et al. [4] dataset are processed into binary values at a 200nt resolution based on a Poisson background distribution. See chapter 2. Nine LRCs are trained in which the covariate was simply the binary value of each histone modification. Moreover, three additional models are trained in which the covariates are combinations of common histone modifications, including linear combination of all nine modifications.

The performance of each LRC, pertaining to individual HMs, was assessed by ROC analysis. A ROC curve plots the number of false positive predictions over the x-axis and the number of correctly predicted positives over the y-axis, at varying threshold values. A common summary statistic in ROC analysis is the so called *area under the curve* (AUC). A model that is perfectly able to discriminate between positive and negative cases (100% sensitivity and 100% specificity) will have an AUC of 1. On the other extreme, a random classifier will have an expected AUC of 0.5. Models with AUC values < 0.5 are negative predictors, and are rare. Moreover, AUC values < 0.5 are often a result of mislabeled positive and negative classifications. For a full summary of ROC analysis, see chapter 6.

Model results (Table 4.1) show that the models, in which the covariate are combinations of histone modifications, perform better than models corresponding to single histone modifications. In fact, Table 4.1 suggests all single HMs except for H3K4me1 are poor predictors, characterized by an AUC of ≤ 0.50. The specificity

		Resampling Results (CV: 10fold, R = 1)		
	Covariate Vector	**AUC**	**Specificity**	**Sensitivity**
1	H3K18Ac	0.3400359	1	0
2	H3K27me3	0.4755401	1	0
3	H3K36me3	0.4923614	1	0
4	H3K4me1	0.7825584	1	0
5	H3K4me2	0.3585582	1	0
6	H3K4me3	0.4012219	1	0
7	H3K9Ac	0.3964052	1	0
8	H4K12Ac	0.4078908	1	0
9	PolII	0.4224162	1	0
10	Methylations only	0.8148944	1	0
11	Acytelations only	0.350107	1	0
12	H3K27me3 + H3K36me3 + H3K4me3	0.377481	1	0
13	All nine histone modifications	0.8349595	0.9996	0.006865

Table 4.1: AUC, Specificity and Sensitivity values for 13 different LRCs corresponding to 13 different covariate vectors. The sensitivity and specificity columns are calculated using a threshold value of 0.5. The confidence intervals are omitted.

and sensitivity columns are calculated using a threshold value of 0.5. Ostensibly, this is insignificant as threshold values are quite arbitrary. In other words, a model's threshold value for which a test result is labeled *positive* need not be 0.5.

As expected, the combination of various HMs have better predictive power [25]. Consider, for example the single modification H3K36me3 which is a poor predictor ($AUC = 0.49$), however is significant in model 11 and model 13. This may be because H3K36me3 is a repressive modification when in promoter regions, but an active modification in the coding region [57]. To this end, it is clear that the combination of all nine HMs have far better predictive power, characterized by a AUC of 0.8653 ($95\%CI : 0.8505 - 0.8838$). In particular all modifications, except H3K4me3, are significant at a 0.05 significance level. The complete specification of this model, including significant coefficients and deviance, is given in Table 4.2. The ROC curves are plotted in Figure 4.1.

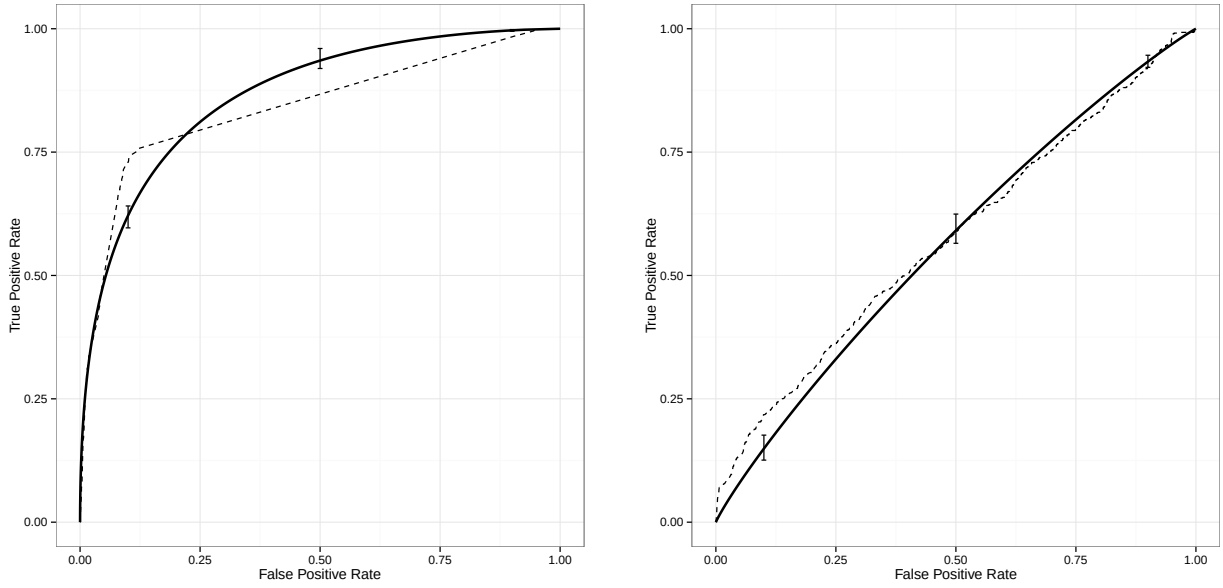
ROC analysis of PWM model In order to accurately evaluate the power of how well chromatin structure predicts TFBS, a systematic comparison is made against the sequenced based PWM model. Recall that the PWM score of a sequence is interpreted to be the probability of achieving a particular score with respect to the background distribution. Notably, the MEF2 specific PWM, constructed in chapter 3, is applied to the

Deviance Residuals				
Min	**1Q**	**Median**	**3Q**	**Max**
-1.3158	-0.1263	-0.1263	-0.1263	3.6963
Coefficients				
	Estimate	Std. Error	Z Value	Pr(>—z—)
(Intercept)	-4.82714	0.03814	-126.561	<2e-16 ***
H3K18Ac	0.33914	0.06123	5.539	3.04e-08 ***
H3K27me3	-1.59802	0.25368	-6.299	2.99e-10 ***
H3K36me3	-1.247	0.16161	-7.716	1.20e-14 ***
H3K4me1	2.33952	0.05764	40.589	<2e-16 ***
H3K4me2	1.10505	0.09268	11.923	<2e-16 ***
H3K4me3	-0.01936	0.09612	-0.201	0.84034
H3K9Ac	0.31393	0.10126	3.1	0.00193 **
H4K12Ac	-0.40509	0.09671	-4.189	2.81e-05 ***
PolII	1.04908	0.08101	12.95	<2e-16 ***

Table 4.2: Estimated parameters and inference statistics for the model in which the covariate was a combination of all nine histone modifications.

same test dataset (see Methods). For all 200nt wide intervals, the PWM score of each interval is defined to be the maximum score of all the subsequences in that interval. Since the test dataset is established from the biological gold standard [62], the true location of MEF2 binding sites are known. Therefore, as before, we employ ROC analysis in order to assess the PWM model. The ROC curves are plotted in Figure 4.1. In contrast to the LRC (AUC 0.8653 ($95\%CI : 0.8505 - 0.8838$)), the PWM model achieves a mediocre AUC of 0.565 ($95\%CI : 0.5551 - 0.584$). This is, of course, expected from chapter 3 and moreover in literature [55]. *In vivo* gene expression is regulated by a multitude of factors, leading us to believe that combining sequence based PWM models with structural HM models can result in better predictive power.

Hidden Markov Models offer a systematic way to analyze histone modifications It is well documented that histone modifications play a crucial role in gene regulation, and hence can be utilized to improve already existing TFBS predictive models. As seen above, the combination of nine histone modifications already have a high predictive power of detecting binding sites. The *histone code hypothesis* suggests that the combinatorial interactions of histone modifications play a great role in gene regulation [57, 9], and thus offers biological significance.



(a) LRC13

(b) Position Weight Matrix

Figure 4.1: (a) Empirical: dashed, $AUC = 0.8413$ $(95\%CI : 0.8505 - 0.8838)$. Smoothed: solid, $AUC = 0.8653$ $(95\%CI : 0.8505 - 0.8838)$) (b) Empirical: dashed, $AUC = 0.5759$ $(95\%CI : 0.5526 - 0.5991)$. Smoothed: solid, $AUC = 0.565$ $(95\%CI : 0.5551 - 0.584)$
Confidence bands for TPR are plotted at $FPR = (0.10, 0.50, 0.90)$. The confidence band for AUC is calculated as defined by Delong *et al.* (1998).

A general and popular framework for modeling histone modification patterns is the Hidden Markov Model (HMM) [24, 41]. An HMM is fully specified by a *Transition Probability matrix* and an *Emission Probability Matrix*, which models local chromatin modification patterns and classifies them into distinct *chromatin states*. The EPM captures the frequency with which different histone combinations are found with each other, and their combinatorial interactions. The entries of the EPM are, therefore, probabilities of interactions associated with each chromatin state. The TPM captures the spatial relationship, but more importantly offers an intuitive and natural way for biologically annotating chromatin states. In other words, every chromatin state, or a group of states, are assigned biological labeling such as *transcription start site* states, *active* states, *repressed* states, etc. The systematic nature of a HMM offers a clear advantage that it allows the model to be applied genome wide, whereas previous models used raw expression data and thus had restrictive prediction regions, mostly close proximity of genes. For a complete description on using HMMs to capture histone modifications, see chapter 2. We apply similar methods from Ernst and Kellis [24] to capture the combinatorial interactions of nine histone modifications, provided by Asp et al. [4], see chapter 2. The resulting nine state HMM for both differentiated and undifferentiated muscle cells, characterized by the Transition Probability Matrix (TPM) and the Emission Probability Matrix (EPM), is given by figure 2.4.

HMM's state assignment as covariates The Hidden Markov Model in Figure 2.4 captures *chromatin states*, defined to be the combinatorial interaction of histone modifications as well as their spatial relationship. In order to use an HMM as a predictive model, a logistic regression classifier is constructed, in which the covariate is the state assignment of each 200nt interval of the training dataset. As a consequence, the covariate variable is categorical with nine levels and thus the interpretation of estimated coefficients for categorical covariates is inherently different than continuous or binary variables. The fitted model is given by Table 4.3. The null state (state 9) is set to be the reference state, and remaining states are tested against this reference. In particular, for each state, the Wald Test is performed to test the difference between the coefficient of the state and the reference state is different from zero. Note that the insignificance of states 6 and 7 do not imply that the entire variable is meaningless. In fact, the overall significance of the variable is obtained by performing the classical ANOVA test. Figure 4.2 is a heat map of when we change the reference state, over all states. The results from the table suggest that states 2 and 3 as well as states 6 and 8 are not significantly different from each other. This is, of course, expected from the results in chapter 2. From Table 4.3, the coefficient of state 9 is the intercept, and so the true coefficient of state 1 is -4.00855. With state 9 as the reference state, for an interval that is assigned state 1, the probability of being is a binding

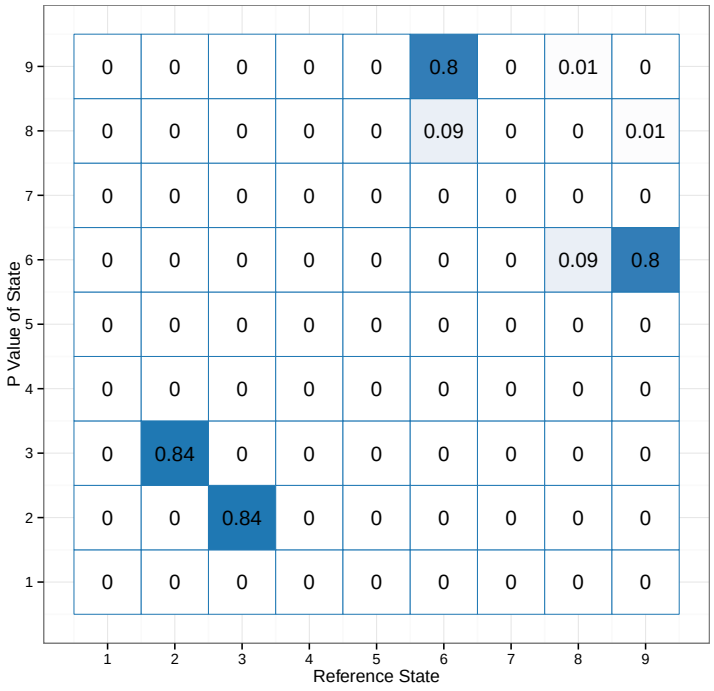


Figure 4.2: This matrix shows the significance of states as the reference modality loops over all the states. For example, accepting the null hypothesis for states 2 and 3 means there is no significant difference between them. Thee highlighted tiles indicate significance (by a Wald Test) at a level of 0.05

site is given by

$$\Pr = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1)$$

where g^{-1} is the inverse logit function. In general, the probability of an interval being a binding site is given by

$$\Pr = g^{-1}(\hat{\beta}_0 + \hat{beta}_1^{s_1} + \beta_2^{s_2} + \hat{\ldots} + \beta_n^{\hat{s}_n})$$

where $\beta_i^{s_i}$ acts as an indicator variable for interval i.

As before, the performance of the model is assesed by ROC analysis. Figure 4.3 plots the ROC curves, having an AUC of 0.8575 with confidence interval CI: $0.8405 - 0.8773$ at a 0.05 significance level. These results are highly supportive of the *histone code hypothesis*. The interactions between the nine histone modifications, embedded in chromatin states of Hidden Markov Model, are highly predictive of *general* transcription factor binding sites.

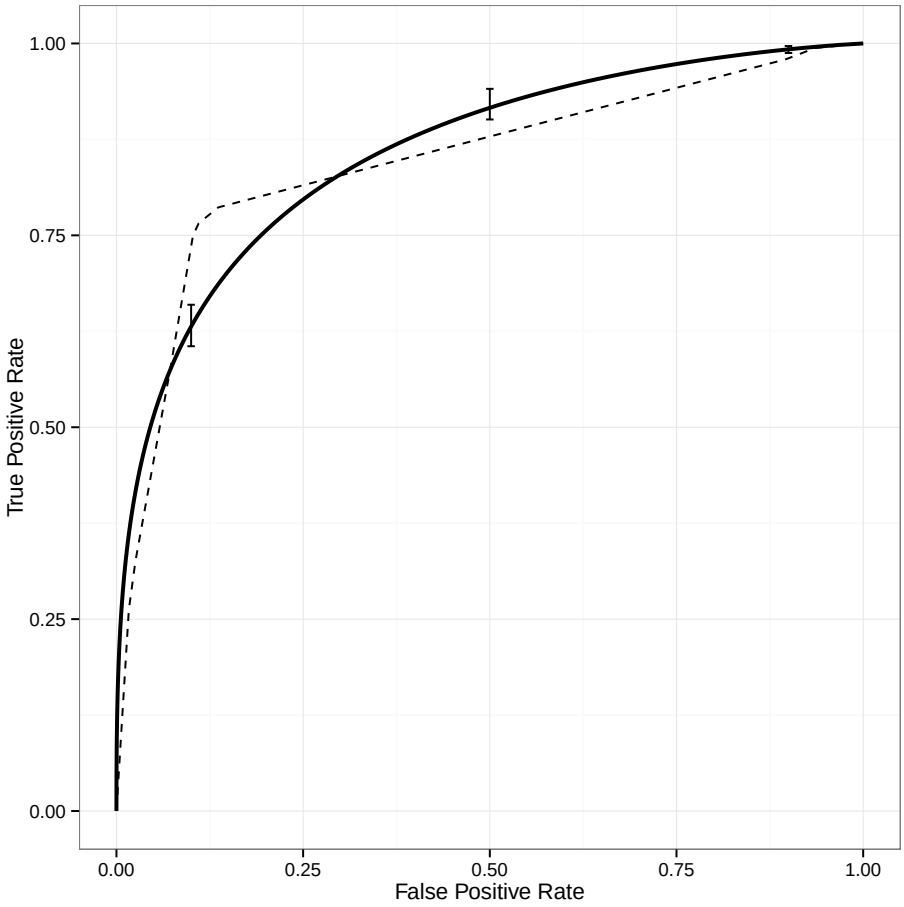


Figure 4.3: The ROC curve of the model in Table 4.3. Confidence intervals for TPR at $0.10, 0.5, 0.90$ are plotted, calculated by Delong's Test.

Deviance Residuals				
Min	**1Q**	**Median**	**3Q**	**Max**
-0.6947	-0.1025	-0.1025	-0.1025	3.4284
Coefficients				
	Estimate	Std. Error	Z Value	Pr(>—z—)
(Intercept)	-5.24658	0.05013	-104.656	<2e-16 ***
State 1	1.23803	0.15855	7.808	5.79e-15 ***
State 2	3.03179	0.07568	40.058	<2e-16 ***
State 3	3.04674	0.06727	45.29	<2e-16 ***
State 4	3.94788	0.07246	54.483	<2e-16 ***
State 5	3.44354	0.10216	33.707	<2e-16 ***
State 6	-0.06039	0.23538	-0.257	0.7975
State 7	2.37798	0.15518	15.324	<2e-16 ***
State 8	-0.62753	0.24798	-2.531	0.0114 *

Table 4.3: Estimated parameters and inference statistics when the covariate is the categorical state assignment value. The z and p values are obtained from the Wald Test (See methods).

TF sequence data are statistically redundant for predicting binding sites The above analysis is focused on how informative HM features are of transcription factor binding sites. The results show conclusively that single histone modifications do not posses sufficient predictive power, however the combinatorial interactions between HMs have regulatory roles that are well established. It can be conjectured that integrating sequenced based MEF2 PWM models with general HM LRC models can generate a highly predictive model for MEF2 binding sites.

To this end, *integrative scores* of 200nt intervals were computed as the product of the PWM score per interval, R, and LRC score per interval G. The integrative scores are simply the product of the two individual model (PWM and LRC) scores. The first set of integrative scores, $R \times G_i$, where $i = 1 \ldots 9$ referring to each of the single HM LRC. An additional model was affixed in which $R \times G_H$ where G_H denotes the score of the LRC corresponding to the nine HM LRC. Ultimately, the last integrative score was $R \times G_S$ where G_S denotes the score of the LRC corresponding to chromatin state LRC.

The first set of single HM integrative scores enlightens if there is a single HM that is particular to MEF2 binding sites. The results in Table 4.4 show an across the board improvement of predictive power of single HM when combined with a sequence model. We further investigated the integrative model when the LRC

	Single HM G_i		$G_i \times$ R	
	AUC	95% CI (Delong)	AUC	95% CI (Delong)
H3K18Ac	0.6696	0.6512-0.688	0.7346	0.7219-0.7415
H3K27me3	0.526	0.5229-0.5291	0.5631	0.5593-0.5816
H3K36me3	0.5051	0.4987-0.5116	0.5705	0.5589-0.5884
H3K4me1	0.7908	0.7731-0.8084	0.8053	0.7958-0.8134
H3K4me2	0.6554	0.6377-0.6731	0.7349	0.7197-0.7509
H3K4me3	0.6947	0.6825-0.7117	0.6169	0.6004-0.6334
H3K9Ac	0.6093	0.5933-0.6253	0.6926	0.6697-0.7122
H4K12Ac	0.5949	0.5794-0.6104	0.681	0.6637-0.7034
PolII	0.5762	0.562-0.5903	0.6615	0.6394-0.6801
All nine histone modifications	0.8653	0.8572-0.876	0.8525	0.8348-0.858

Table 4.4: AUC values of single histone LRCs and the integrative score model.

was trained with the combination of all nine histone modifications. The integrative model (Figure 4.5 had an AUC value of 0.8525, whereas the HM only model had an AUC of 0.8653. The almost-equal AUC values suggest that the PWM does not offer an improvement over the predictive power. This may be due to the low information content of the PWM as the PWM is constructed based on different MEF2 isoforms, each with distinct consensus sequence. Previous studies have shown that individual transcription factors are statistically redundant for predicting gene expression when the number of histone modifications exceeded four [9].

Ultimately, the integrative scores in which the LRC is constructed from chromatin states is of most interest. Recall that the chromatin states capture the combinatorial interactions between HMs. Further recall that the chromatin state LRC had an AUC value of 0.8575. In contrast, the integrative model has AUC of 0.8607 which is only marginally better. The ROC curves are plotted in Figure 4.5. Admittingly, this is not expected as it was conjectured that the integrative models would improve accuracy. However, there is sufficient evidence that integrative models do not always provide better predictive power. Budden et al. [9] shows that transcription factors and histone modifications provide equivalent information in genome-wide gene regulation.

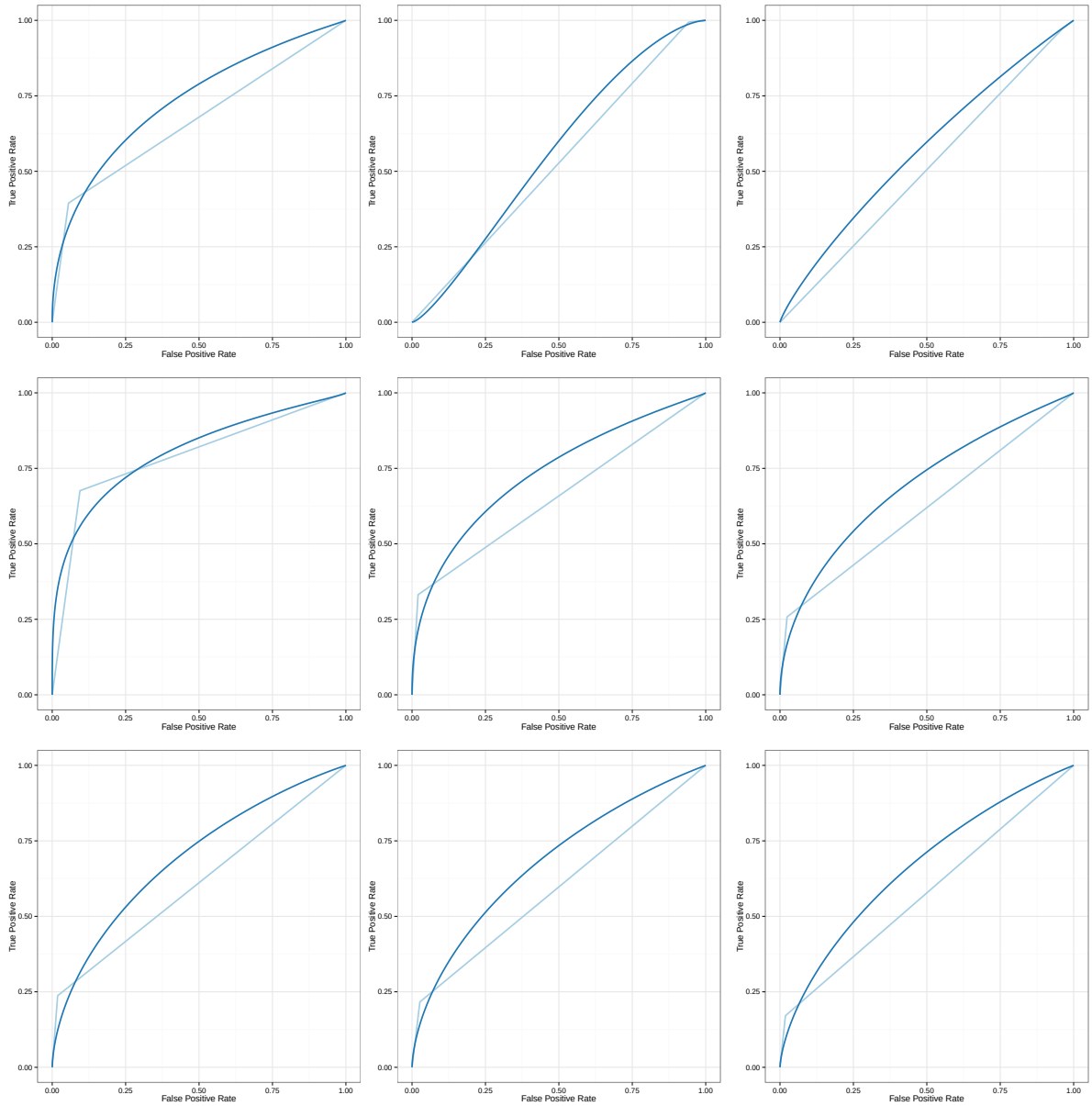


Figure 4.4: ROC curves for every single HM LRCs and the PWM integrative models. The curves for the integrative model was smoothed using methods described in chapter 6. The AUC values and confidence bands are given in Table 4.4

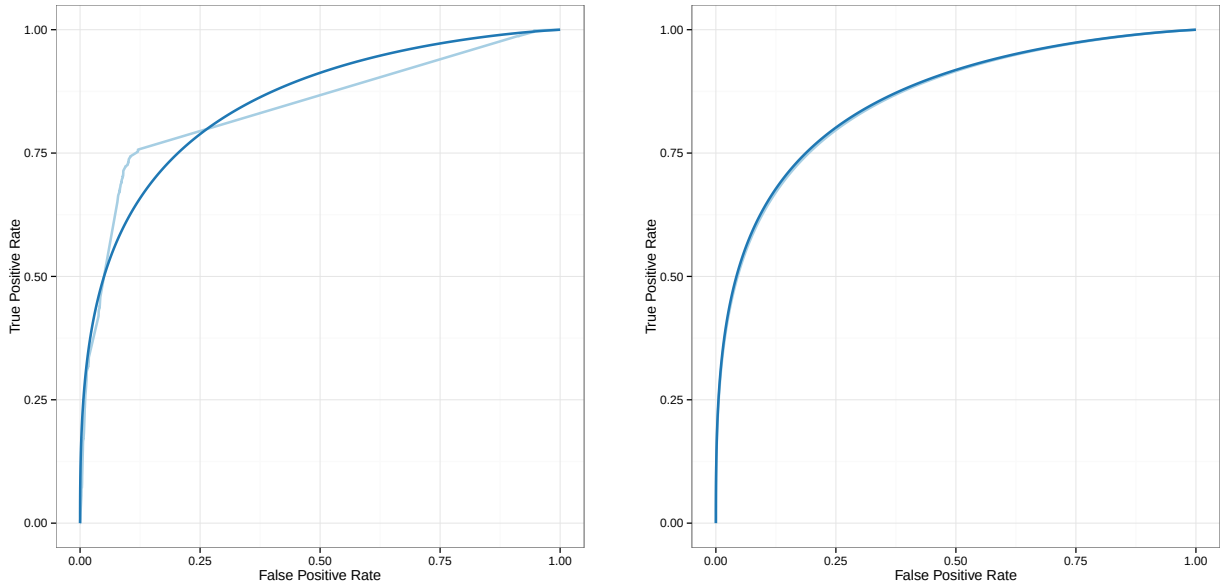


Figure 4.5: (a) The integrative model combining PWM scores and the 9 HM LRC model (b) The integrative model combining PWM scores and the chromatin state LRC model.

4.4 Discussion

As TFs and HMs both play critical roles in gene regulation, accurate predictive models can be constructed by integrating the individual datatracks. In this chapter, we have successfully probed factors that are predictive of transcription factor binding sites, by integrating histone modifications with sequence motif data. The first set of models were constructed on information provided by single HMs but also interactions between histone modifications. The different combinatorial interactions between HMs were captured into nine distinct chromatin states, using a Hidden Markov Model. We expect that HMs are informative in predicting binding sites. Indeed, distinct chromatin states are highly predictive of transcription factor binding sites, characterized by a high AUC value. We find that individual HMs are *equally* predictive of binding sites, however do so weakly. This suggests that individual histone modifications are statistically redundant as predictive sources. This may be due to the fact that histone modifications are closely correlated and there is informative redundancy between them. It is important to note that that HM-based models provide accurate predictions for binding of *any* transcription factor.

In chapter 3, we constructed a PWM model specific to the transcription factor MEF2. The PWM did not show to have sufficient accuracy with an AUC value of 0.65. We hypothesized that the integration of the

HM model with the PWM model would offer better prediction accuracy. To investigate this hypothesis, we constructed an integrative model and applied it to the test set. In particular, the score of each 200nt interval of the test set was given by the product of the HM score and the PWM score of the interval. As expected, comparing the ROC curves and the AUC values, this new integrative model is highly predictive of MEF2 binding sites.

Although TF and HMs are informative sources for predicting binding sites, it is important to note that the AUC difference between the integrative model and the chromatin state HM model is negligible. This suggests that the information provided by the PWM model is statistically redundant and not significant. The HM-only model trained with chromatin state data had an AUC of 0.8575. In contrast, the HM-PWM method, had an AUC of 0.8607. It is apparent that these results contradict what we were expecting. However, this is line with previous studies. The study by Budden et al. [9] concludes that transcription factors and histone modifications provide equivalent information regarding genome-wide gene regulation. It is also important to note that the various assumptions used in constructing the PWM model may have lead to these results.

4.5 Methods

Training and Test Datasets Genome wide coordinates of the MEF2 transcription factor were obtained from Wales et al. [62]. There are 2797 peaks identified in their experiment. Due to the nature of ChIP-EXO experiments, not every single binding site is captured. Furthermore, MEF2 has four isoforms: MEF2A, MEF2B, MEF2C, and MEF2D. The experiment by Wales et al. [62] is mainly focused on the MEF2A isoform. Each isoform has a different consensus sequence; an important point to consider when using PWM models. Lastly, the number of binding events depend on the tissue, time, and cell conditions of the ChIP-EXO experiment.

The positive training dataset was selected as follows: For each true binding site, as reported by Wales et al. [62], the bin number in which the binding site lies was calculated. The bin resolution of 200nt is sufficient as each nucleosome is covered by DNA that is 147nt long. In other words, the binding site which most likely exists on a nucleosome is affected by the modified histones on the nucleosome. For the negative training set, we randomly sampled 49 bins (known to not have a binding site) for every positive bin. There were no restrictions placed on where these 49 bins came from, however as suggested by Ernst et al. [25], it might be

beneficial to only select bins that come from non-gapped regions of the genome. We performed stratified random sampling so that for every one positive bin on a chromosome, we get 49 negative bins from the same chromosome. This sets a prior expectation that, on average, 2% of the genome is bound by transcription factors. This figure is biologically intuitive as well, since 3.5% of the genome is believed to have non-protein coding functionality [25].

Logistic Regression Classifier We used the statistical framework of generalized linear models to integrate histone modification data, PWM scores, and HMM. Let y_i be a binary response such that $y_i = 1$ if the i'th interval on a sequence contains a binding site, $y_i = 0$ otherwise. The binary response variable can be interpreted as a realization of a binary random variable Y_i, with $\mathrm{E}[Y_i] = \pi_i$. The mean π_i depends on a vector of observed covariates $\mathbf{x}_i$. The covariates can be either categorical, ordinal, or continuous data. In the context of this chapter, we use both categorical covariates (HMM states) and continuous covariates (raw HM data). Since the covariates are real valued and $0 \leq \pi_i \leq 1$, a transformation of π_i is required to remove the range restriction. This leads to utilizing the *logit* function for the transformation, and the so called *link function*. The Logistic Regression Classifier model is given by

$$\eta_i = \boldsymbol{\beta}\boldsymbol{x_i}^T \tag{4.1}$$

where $\boldsymbol{x}$ is a vector of the covariates and $\boldsymbol{\beta}$ is a vector of the regression coefficients. The model specified in (4.1) is a generalized linear model, with a binomial response variable and a sigmoid link function so that $g(\pi_i) = \eta_i$. Applying estimating methods to solve for $\boldsymbol{\beta}$, the function

$$h(\boldsymbol{x_i}) = g^{-1}(\boldsymbol{\beta}\boldsymbol{x_i}) = \frac{e^{\boldsymbol{\beta}\boldsymbol{x_i}^T}}{1 + e^{\boldsymbol{\beta}\boldsymbol{x_i}^T}}$$

represents the probability of a region i containing a binding site, ie

$$h(\boldsymbol{x_i}^T) = \Pr(y_i = 1 \mid \boldsymbol{x_i}$$

Position Weight Matrix Specification A Position Weight Matrix (PWM) is a probabilistic model to identify potential binding sites in a target sequence. A zero order PWM given by table 3.2 in chapter 3, applied to any sequence of length L, returns $w - L + 1$ scores corresponding to the $w - L + 1$ subsequences.

These scores represent the binding affinity of the TF to the sequence. In this context, the PWM is applied to all 191 subsequences denoted u_k, of a bin i. We, then, define the overall score for bin i by

$$R(i) = \max_{k=1-191}[R_+(u_k), R_-(u_k)] \tag{4.2}$$

where R_+, R_- denote the scores on the positive and negative sequence strands.

4.6 Literature Review

Previous studies [9, 25, 57, 12, 44] have confirmed that both TF and HMs play a crucial role in predicting gene expression. These studies have developed models in which the information of HM have come from raw gene expression data. Hence, the accuracy of these models rely on experimental wet-lab datasets which are highly tissue and cell condition specific. Nonetheless, some of these models have suggested that TFs and HMs are more accurate predictors for transcription factor binding sites than using PWM models alone [9].

Recently, a computational model proposed by [57] to improve binding site discovery by considering nucleosome positioning. They discovered that using the genomic positioning of modified nucleosome can be informative for predicting transcription factor binding sites. Their first approach, *Modified Nucleosome Neighboring*, showed that the vicinity of modified nucleosomes around TF binding sites combined with PWM scores improves the false discovery rate over using the PWM alone. As a consequence of this approach, the study found that seven particular histone modifications (H3K4me1, H3K4me2, H3K4me3, H4K20me1, H2BK5me1, H3K9me1, and H3K27me1) are high correlated with transcription factor binding sites. The study used this information to develop a secondary approach, *Modified Nucleosome Occupancy*, to analyze the frequency of modifications around TFBSs. Their methods utilized the logistic regression classifier (LRC) with the sigmoid function

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

to integrate the data sources.

Similarly, the model by Cuellar-Partida et al. [12] develops a novel heuristic method to integrate a prior distribution from epigenetic data. Their results also suggest that that histone modifications H3K4me1, H3K4me3, H3K9Ac, and DNase1 sensitivity conclusively improve TFBS prediction over a PWM model.

Notably, their algorithm is now embedded in the popular bioinformatics *MEME Suite* toolkit. A large scale, genome wide study by Ernst et al. [25] incorporates 29 additional information sources, including distance from the nearest TSS, levels of histone modification, CpG islands, and evolutionary traits. In particular, they used 20 different histone modifications and concluded that combining these information sources improves prediction of TF bound regions.

A study by McLeay et al. [44] successfully explained gene expression patterns by building an integrated model of 12 transcription factors, several histone modifications and DNase hypersensitivity. Similar to our results, the study concludes that the seven histone modifications as well as DNase data can explain up to 70% of the variance in gene expression in mES cells. Furthermore, the study found evidence that models in which histone modification data was combined with TF ChIP-seq data performed better than models constructed with TF ChIP-seq data alone.

5 A summary of Generalized Linear Models and Logistic Regression

5.1 Introduction

In the interest of keeping this thesis self-contained, this chapter hopes to provide a brief exposition on methods for statistical modeling. In general, mathematical modeling is where one establishes a method, or *trains a model*, to explain variation in data and further use the model to draw predictions. Statistical techniques and principles are applied so that the *trained* models are well suited for **predicting** an outcome, given some explanatory variables. Formally, the explanatory variables (independent variables) are non-random measurements or observations. A quantitative explanatory variable is called a covariate. The response variable (dependent variable) are free to vary in response to the explanatory variables. Common statistical models are those for which several explanatory variables decide a single response variable. The general work flow in mathematical modeling include *formulating*, *estimating*, *validating*, and *testing* models for the main purpose of *predicting* the mean value of random variables. Different types of data require different modeling methodologies. In practice, several types of *response* variables are seen such as

- *Continuous Data*($y_1 = 5.4, y_2 = 9.2, y_3 = -1.2, \ldots, y_n = 0.9$). Examples of this type of data include air temperature and precipitation. This data often follows a normal distribution. A special case of continuous data is *Continuous Positive Data* in which, as the name implies, the response variable is greater than zero. This type of data of comes when dealing with concentrations and is log-normally distributed.

- *Count Data*($y_1 = 5, y_2 = 10, y_3 = 0, \ldots, y_n = 12$). Examples of this type of data include car accidents and the number of customers walking into a store. Poisson distributed.

- *Binary Data*($y_i = 0$ or 1). Examples include admittance and rejections from universities. Binomial distributed.

- *Nominal Data* or alternatively, categorical data. This type of data can be unordered (e.g. Male/Female) or ordered (Rating from 1 to 5). Multinomial distribution.

The methods explained in this chapter are focused primarily on binary data, which is well modeled by a Generalized Linear Model. It is assumed that the reader knows basic statistics including common distributions such as Binomial and their properties. To that end, only basic theory and principles of Generalized Linear Models (GLM) are provided. GLM is a modeling framework when considering data (observations, response) that follows the so called *exponential family of distributions*. The methods contrast with *General Linear Models* which are relevant only for Gaussian (Normally) distributed data. In a general linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \epsilon_i$$

, the response variable $y_i, i = 1 \ldots n$ is modeled by a linear function of the explanatory variables $x_i, i = 1 \ldots n$ plus some error term. The linearity of the model is in the parameters β_i. The errors are independent and identically distributed such that $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$, ie $\epsilon_i \sim N(0, \sigma^2)$ as a basis for inference. Although, the general linear model is useful, it is not appropriate when the response is binary or count data. In general, when the range of the response variable Y is restricted or the variance of Y depends on the mean, general linear models are inefficient models. Generalized Linear Models, are thus, and extension of general linear models and address the issues above.

5.2 Generalized Linear Models

.

Exponential Family of Distributions Most of the commonly used statistical distributions (Normal, Binomial, and Poisson) belong to the family of exponential distributions [18]. A random variable Y that belongs in this family has a density function written in the form

$$f_Y(y, \theta) = c(y, \lambda) \exp(\lambda()\theta y - \kappa(\theta))), \quad \theta \in \Omega \tag{5.1}$$

Here, Ω is the parameter space. The function $\kappa(\theta)$ is called the cumulant generator. The parameter θ is called the *canonical parameter*, parameter λ is called the *precision parameter*. Many of the properties of distributions that can be written in the form 5.1 can be derived from the cumulant generator. In particular, if a RV Y has a distribution in the form of 5.1, then

$$\mathrm{E}[Y] = \kappa'(\theta) \tag{5.2}$$

$$\mathrm{E}[Y] = \frac{\kappa''(\theta)}{\lambda} \tag{5.3}$$

Note that the function $\tau(\theta) = \kappa'(\theta)$ defines a one-to-one mapping of the parameter space Ω onto a subset S of the real line. This subset is called the *mean value space*. The mean value space can be roughly though of as the convex hull of the support of the distribution [43]. The inverse mapping $\theta = \tau^{-1}(\mu)$ is called the *canonical link function*.

An Integrative Model Overview The generalized linear model is defined in terms of a set of n independent random variables $Y_i, i = 1 \dots n$, representing the response y_i, each with a distribution that belongs to the exponential family. This set of random variables satisfies properties 1) the distribution of each Y_i depends only on the parameter θ_i and 2) the distributions of all Y_i are the same, thus have the same cumulant generator $\kappa(\cdot)$. The joint density is then given by

$$f(y_1, \dots, y_n; \theta_1, \dots, \theta_n) = \exp\left[\sum_{i=1}^{n} \lambda_i(\theta_i y_i - \kappa(\theta_i))\right] \prod_{i=1}^{n} c(y_i, \lambda_i) \tag{5.4}$$

The parameters θ_i are generally not of interest [18]. For a generalized linear model, we are usually interested in estimating a smaller set of parameters $\beta_1, \beta_2, \dots, \beta_p$ (where $p < n$). This set of β_i's is such that the linear combination of them (the **linear predictor**) is equal to some function of the expected value μ_i of Y_i, ie

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \beta$$

The *linear predictor* η describes a function of the mean value and incorporates information about the independent variables into the model. The *link function* $g(\cdot)$ provides the relationship between the linear

predictor η and the mean value parameter $\mu = \mathrm{E}[Y]$. In particular, we have

$$\eta = g(\mu)$$

The inverse mapping $g^{-1}(\cdot)$, therefore, describes the mean value as a function of the linear predictor, ie

$$\mu = g^{-1}(\eta)$$

There are many choices for the link function, and the choice is somewhat arbitrary [43]. In summary, a generalized linear model has three components:

1. Independent and identically distributed random variables for the response variables

$$Y_1, Y_2, \ldots, Y_n$$

2. A set of parameters $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_n]$ and independent (explanatory) variables $\mathbf{x} = [x_1, x_2, \ldots, x_n]$

3. A monotone link function g such that

$$g(\mu_i) = \mathbf{x_i^T}\boldsymbol{\beta}$$

where $\mu_i = \mathrm{E}[Y_i]$.

5.3 Model Estimation

Estimation of the model parameters $\boldsymbol{\beta}$ can be done by Maximum Likelihood methods or Bayesian methods. The log-likelihood for a model specified as above is

$$l = \sum_{i=1}^{n} \frac{y_i\theta_i - \kappa(\theta_i)}{\lambda_i} + c(y_i, \lambda_i) \tag{5.5}$$

The estimates for parameters can be obtained by solving the *score* equations

$$s(\beta_i) = \frac{\partial l}{\partial \beta_i} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\lambda_i V(\mu_i)} \times \frac{x_{ij}}{g'(\mu_i)} = 0 \tag{5.6}$$

where x_{ij} is the j'th element of $\boldsymbol{x_i}$. A nice property of the exponential family of distributions is that they ensure that the global maximum of the log-likelihood function $l(\boldsymbol{\theta}, \boldsymbol{y})$ is given uniquely by the solution of the score equations [18]. A general method of solving the score equations is the iterative weighted least squares by the Newton-Raphson method. The r-th iteration, the new estimate for $\beta^{(r+1)}$ is obtained by

$$\beta^{(r+1)} = \beta^(r) + \mathfrak{J}^{-1}(\beta^{(r)})s(\beta^{(r)}) \tag{5.7}$$

where $\mathfrak{J}$ is the *observed information matrix* (the Hessian Matrix with the opposite sign).

5.4 Logistic Regression

Logistic Regression is utilized in the scenario when the response variable y_i is binary. In this context, y_i is considered a realization of a random variable Y_i with the Bernoulli distribution, and can be written as

$$\Pr(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

for $y_i = \{0, 1\}$. The expected value and variance of Y_i are $\mathrm{E}[Y_i] = \mu_i = \pi_i$ and $\mathrm{var}[Y_i] = \sigma_i^2 = \pi_i(1 - \pi_i)$. Since the mean and variance of Y_i depend on the probability π_i, a linear model is not sufficient as it assumes constant variance. Therefore, the application of GLMs is required.

The Logit transformation In order to systematically establish the logistic regression model, a relationship is required between the probabilities π_i and the observed covariates $\boldsymbol{x_i}$. The relationship $\pi_i = \boldsymbol{\beta x_i}$ is not sufficient due to the natural range restriction on π_i and the real valued RHS linear predictor. A transformation of the probabilities can be applied to remove the range restriction. In particular,

$$\eta_i = \mathrm{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}$$

The *logit* is a one-to-one transformation that maps probabilities in $(0, 1)$ to R. By GLM terminology, the logit function is precisely the *link* function. The inverse link function allows to go back to probabilities

$$\pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

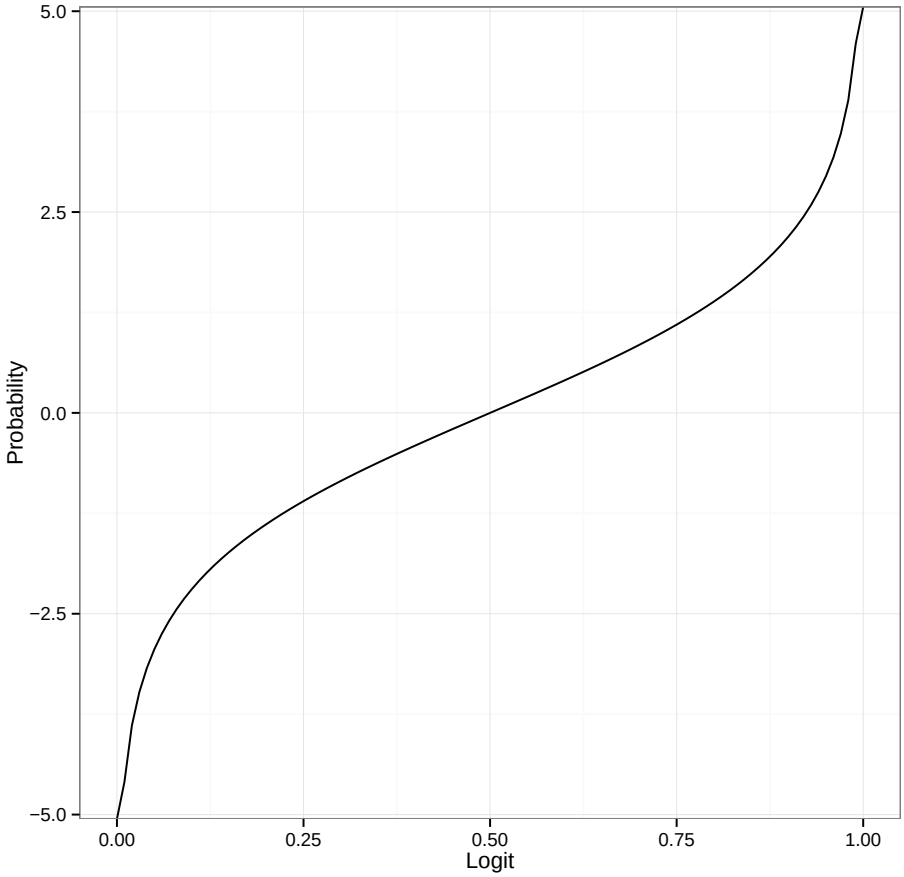


Figure 5.1: The logit function.

Specification of the model The structure of a logistic regression model is defined by the random variable $Y_i \sim B(n_i, \pi_i)$ where i ranges from 1 to k different, distinct observations. Formally, the *logistic regression model* is

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \boldsymbol{x_i}\boldsymbol{\beta}$$

Estimation methods are applied to retrieve the coefficients $\boldsymbol{\beta}$. The inverse link function yields the probabilities again. Suppose, $Y_i = 1$ (positive) when $\pi_i \geq 0.5$ and $Y_i = 0$ (negative) when $\pi_i < 0.5$. As a consequence the logistic regression model is equivalent to a linear classifier [18]. In general, the decision boundary separating positive and negative classes is given by the solution to $\boldsymbol{x}\boldsymbol{\beta} = 0$. Those familiar with linear algebra will notice that the decision boundary is a point if x is one-dimension, a line if x is two-dimensional, a place if x is three-dimensional, and so on.

6 A summary of Receiver Operating Characteristics

6.1 Introduction

In the interest of keeping this thesis self-contained, this chapter hopes to provide a brief exposition on methods for assessing the performance of binary classifiers. The Receiver Operating Characteristic (ROC) curve is one of the best developed statistical tool to evaluate binary classifiers. ROC curves have gained tremendous popularity since its development by World War II engineers for signal detection theory. Its utilization has quickly expanded into many other fields including biosciences, psychology, finance and sociology. In particular, it is widely applied in medicine to evaluate diagnostic tests discriminate diseased from normal cases [47]. For instance, radioactive imaging is common diagnostic test in which the test results are real numbers. The higher (or lower) continuous value of the test indicates the presence (or absence) of a disease. Applying ROC analysis evaluates the discriminatory ability of the radioactive imaging diagnostic test, assuming the true status of the disease is known. In general, ROC analysis returns a measure of the discriminatory ability of any continuous, two-group classifier (true/false, yes/no, positive/negative, diseased/non-diseased) as long as the true status of the cases are known by an independent means of testing. This chapter provides an overview on some inference and estimation methods for constructing ROC curves and its associated summary measures.

The object of interest in ROC analysis is the so called *ROC curve* which is a graphical representation of the relationship between the false positive rate and the true positive rate of any classifier. The true positive rate, also known as sensitivity, of a classifier is the probability that a `TRUE` object is correctly classified by the model. Similarly, the specificity Sp is the probability that a `FALSE` object is correctly rejected by the model and the false positive rate is, therefore, $1 - \text{Sp}$. In the context of medical diagnostics, Se represents the probability that a truly diseased individual has a positive test result and Sp is the probability that a

truly non-diseased individual has a negative result. The ROC curve characterizes Se as a function of $1 - \text{Sp}$. In other words, the ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR), for various threshold values. The sensitivity and specificity rates allow us to rigorously analyze the classifier by using conditional probabilities of belonging to a particular class given the true classification. In statisticalterms,thesecurvesdisplaythe trade-ofbfetweenpowerandsize ofthetestwithrejection regionsX ¿ 0 asthethreshold0isvaried

The area under the ROC curve (AUC) has been established as a fundamental summary measure of a classifier's accuracy. The AUC is interpreted as the probability of correctly classifying between a randomly selected pair of `TRUE` and `FALSE` objects. More intuitively, given a randomly selected pair of nondiseased and diseased individuals, the classifier assigns a higher score for the diseased subject. AUC values close to 1 suggest an almost perfect classifier. On the other hand, values close to 0.5 suggest an essentially useless classifier. In other words, an area of 0.5 suggests that the diagnostic test was only able to classify 50% of the cases correctly. This is no better, essentially, flipping a coin.

For the rest of this chapter, we present a few theoretical results of ROC analysis and describe methods for creating ROC curves. The termininolgy used for the rest of the chapter is in the context of a medical test. To this end, the binary classifier is some diagnostic test which returns a continuous result. The populations are continuous random variables grouped into non-diseased (X) and diseased (Y) with size $n_N + n_D$, respectively.

6.2 Background

Let $X \sim F$ and $Y \sim G$ be two continuous random variables representing two populations: non-diseased and diseased respectively. Let c_t be a threshold value, such that a patient is classified as sick if the diagnostic test score is greater than c_t. We borrow the notation of Pepe [47] and [11] for what follows. For a given threshold $c_t \in \mathbb{R}$, we define the false positive and true positive rates as

$$FP(c_t) = \Pr(X > c_t) = \int_{-\infty}^{\infty} f_X(x) I(x - c_t)\,\mathrm{d}x \tag{6.1}$$

$$TP(c_t) = \Pr(Y > c_t) = \int_{-\infty}^{\infty} g_Y(y) I(y - c_t)\,\mathrm{d}y \tag{6.2}$$

where I is the indicator function. The ROC curve, which plots TP rate against FP rate, is obtained by

$$(t, R(t)) = (FP(c_t), TP(c_t)) \quad c_t \in \mathbb{R}$$

where $t \in [0, 1]$. When the false positive rate t is given, then

$$\begin{aligned} t = FP(c_t) &= \Pr(X > c_t) = 1 - F(c_t) = \bar{F}(c_t) \\ \implies c_t &= [1 - F(c_t)]^{-1} = F^{-1}(1 - t) \end{aligned}$$

where $\bar{\cdot}$ are the survival functions, $F^{-1}(\eta) = \inf(x \mid F(x) > \eta)$ and the relation $[1 - F(x)]^{-1} = F^{-1}(1 - x)$ by setting $\pi(x) = 1 - x$ and using the general identity $(\pi \circ F)^{-1} = F^{-1} \circ \pi^{-1}$. Therefore if $F^{-1}(1 - t)$ exists, the functional form of the ROC curve is given by

$$R(t) = TP(c_t) = \Pr(Y > c_t) = 1 - G(c_t) = 1 - G(F^{-1}(1 - t)) \tag{6.3}$$

In statistical analysis, $R(t)$ represents the distribution function for testing the null hypothesis that the individual being tested comes from the non-diseased population. It is easy to see that as c_t increases, both $TP(c_t)$ and $FP(c_t)$ decrease. Particularly, when $c_t = \infty$, we have $\lim_{c_t \to \infty} TP(c_t) = 0$ and $\lim_{c_t \to \infty} FP(c_t) = 0$. On the other hand, when $c_t = -\infty$, we have $\lim_{c_t \to \infty} TP(c_t) = 1$ and $\lim_{c_t \to \infty} FP(c_t) = 1$. Thus, the ROC curve is a monotone increasing function that maps $(0, 1)$ onto $(0, 1)$. Any diagnostic test is as good as a random classifier if $R(t) = t$, the unit slope line. In this case, the test is essentially useless (or no better than flipping a coin). A perfect test, on the other hand, can fully discriminate between diseased and non-diseased subjects. That is, for some threshold c_t, we have $TP(c_t) = 1$ and $FP(c_t) = 0$.

Area Under the ROC Curve The extensively used summary measure, AUC, is numerical value used to the convey important information about the curve. It is defined and estimated by

$$\text{AUC} = \int_0^1 R(t)\mathrm{d}t \quad \text{and } \hat{\text{AUC}} = \int_0^1 \hat{R}(t)\mathrm{d}t \tag{6.4}$$

A diagnostic test that can fully discriminate between diseased and non-diseased subjects (ie. a perfect classifier) has area $AUC = 1$. Conversely, a random (useless) classifier, $R(t) = t$, has $AUC = 0.5$. The area under a ROC curve is interpreted as the probability of correctly classifying between a randomly selected pair

of diseased and non-diseased subjects, ie AUC $\sim \Pr(Y > X)$. To see this, recall that $X \sim F$ and $Y \sim G$ are the continuous random variables for non-diseased and diseased subjects. By Equation 6.3 and Equation 6.4, we have

$$\begin{aligned} A = \int_0^1 R(t)\,\mathrm{d}t &= \int_0^1 1 - G(F^{-1}(1-t)) \\ &= \int_0^1 \bar{G}(\bar{F}^{-1}(t))\,\mathrm{d}t \end{aligned}$$

Let $y = \bar{F}^{-1}(t)$ so $\bar{F}(y) = t$

$$\begin{aligned} &= \int_{-\infty}^{\infty} \bar{G}(y)\,\mathrm{d}\bar{F}(y) \\ &= \int_{-\infty}^{\infty} \Pr(Y > y) f_X(y)\,\mathrm{d}y \end{aligned}$$

Since X, Y are independent,

$$\begin{aligned} &= \int_{-\infty}^{\infty} \Pr(Y > y \text{ and } X = y)\,\mathrm{d}y \\ &= \Pr(Y > X) \end{aligned}$$

where $\bar{F}$ and $\bar{G}$ are survival functions for the random variables X and Y. Although, this interpretation of correctly identifying a random pair of diseased and non-diseased subjects is sufficient for this thesis, it is not neccessarily the best interpretation in medical diagnostic tests. Pepe [47] provides the interpretation that the AUC is an average TPR, averaged uniformly over the whole range of false positives in $(0, 1)$. This naturally leads to the idea of *partial* AUC Pepe [47]. By fixing a particular false positive rate, t_0, values of $R(t), t < t_0$ provide significant meaning when values of $t > t_0$ are not of interest. The partial area under the curve $pAUC(t_0)$ is a summary measure that restricts the false positive rate at $\leq t_0$. See Pepe [47] for a full exposition.

6.3 Parametric Method to calculate AUC

Parametric methods are used when the distribution functions F and G, for non-diseased and diseased populations, is known. We use the **binormal method** to provide exposition on parametric ROC analysis. The choice to use the binormal method to estimate ROC curves is usually justified by its mathematical rigor, familiarity the normal distribution or just by convenience. It allows for easy estimation of the curve parameters using the means and variances of the classifier values [47]. The binormal method requires to assume that the diagnostic test scores for both diseased and non-dieseased populations follow normal distributions.

Let $X \sim N(\mu_N, \sigma_N^2)$ and $Y \sim N(\mu_D, \sigma_D^2)$ be independent distributions coming from two populations: non-diseased and diseased. Then by Equation 6.1 we have

$$\begin{aligned} FP(c_t) &= \Pr(X > c_t) \\ &= 1 - \Phi\left(\frac{c_t - \mu_N}{\sigma_N}\right) \\ &= \Phi\left(\frac{\mu_N - c_t}{\sigma_N}\right) \end{aligned}$$

and

$$\begin{aligned} TP(c_t) &= \Pr(Y > c_t) \\ &= 1 - \Phi\left(\frac{c_t - \mu_D}{\sigma_D}\right) \\ &= \Phi\left(\frac{\mu_1 - c_D}{\sigma_D}\right) \end{aligned}$$

For a given false positive rate, t, $c_t = \mu_n - \sigma_N \Phi^{-1}(t)$ is the corresponding threshold for the true positivity. Hence,

$$\begin{aligned} R(t) = TP(c_t) &= \Phi\left(\frac{\mu_D - c_t}{\sigma_D}\right) \\ &= \Phi\left(\frac{\mu_D - \mu_N + \sigma_N \Phi^{-1}(t)}{\sigma_D}\right) \\ &= \Phi(a + b\Phi^{-1}(t)) \end{aligned} \tag{6.5}$$

where $a = \frac{\mu_N - \mu_D}{\sigma_N}$ and $b = \frac{\sigma_N}{\sigma_D}$ is the intercept (*separation*) and slope (*symmetry*) coefficients. The area

A under the ROC curve, representing the probability that a randomly selected diseased subject has a classifier score higher than a randomly selected non-diseased subject, now has a simple analytic form. Since $AUC = \Pr(Y > X)$. Let $W = Y - X$, then

$$W \sim N(\mu_D - \mu_N, \sigma_D^2 + \sigma_N^2)$$

and

$$\begin{aligned}
\Pr(W > 0) &= 1 - \Phi\left(\frac{\mu_D - \mu_N}{\sqrt{\sigma_D^2 + \sigma_N^2}}\right) \\
&= \Phi\left(\frac{\frac{\mu_D - \mu_N}{\sigma_D}}{\sqrt{1 + \frac{\sigma_N^2}{\sigma_D^2}}}\right) \\
&= \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right) \qquad (6.6)
\end{aligned}$$

It is easy to see that the AUC is a monotonic increasing function of a and a decreasing function of b. The estimated parameters a and b (denoted $\hat{a}$ and $\hat{b}$) are computed using $\hat{\mu}$ and $\hat{\sigma}$. These can further be obtained by well established estimation methods such Maximum Likelihood Estimatation and Bayesian Statistics. If using ML methods, the variance/covariance of $\hat{a}, \hat{b}$ can be estimated from Fishers information matrix. Figure 6.1 is an example of a ROC curve constructed by the binormal method.

6.4 Non-Parametric Methods to calculate AUC

More often than not, the distribution of the test scores is not known or may not exhibit normality. The empirical method, a nonparametric approach, is the statistical methadology for making inferences about the ROC curve when the underlying distribution is not known. The empirical estimator of the ROC curve is simplistic method based on plugging in empirical evidence into Equation 6.3. This method is popular since there is no assumption about the underlying distribution of the diagnostic test scores. Let n_N and n_D denote the number of non-diseased and diseased subjects. The corresponding true positive and false positive rates

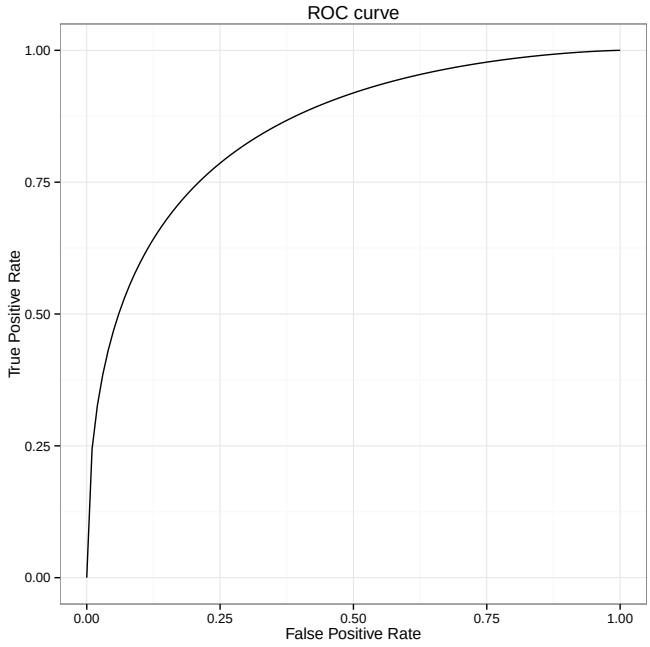


Figure 6.1: Example of a ROC curve for a bi-normal model, constructed using Equation 6.3 with the Normal Distribution $N\left(\frac{a}{b}, \frac{1}{b}\right)$ where $a = 1.4$ and $b = 0.9$.

for every threshold value c are calculated by

$$\mathrm{TP}(c) = \frac{s_D(c)}{n_D} \tag{6.7}$$

$$\mathrm{FP}(c) = \frac{s_N(c)}{n_N} \tag{6.8}$$

where $s_1(c)$ is the number of subjects with test scores greater than c amongst the diseased subjects and $s_0(c)$ is the number of subjects with test scores greater than c amongst the non-diseased subjects. We can write the empirical ROC curve as

$$\hat{R}(t) = \hat{\bar{G}}(\hat{\bar{F}}^{-1}(t)) \tag{6.9}$$

where $\hat{\bar{F}}$ and $\hat{\bar{G}}$ are the empirical survival functions of of X and Y and $\hat{\bar{F}}^{-1}$ is the empirical quantile function. For every value of c, the above equations return a point in the ROC space, and the ROC curve is constructed by joining these points by straight lines. The area under the curve A calculation is provided by the trapezoidal algorithm and is estimated by

$$\hat{A} = \frac{1}{n_N n_D} \sum_{i=1}^{n_0} \sum_{j=1} \psi(Y_{iD}, Y_{jN}) \tag{6.10}$$

where

$$\psi(Y_{iD}, Y_{jN}) = \begin{cases} 1 & \text{if } Y_{iD} > Y_{jN} \\ 1/2 & \text{if } Y_{iD} = Y_{jN} \\ 0 & \text{if } Y_{iD} < Y_{jN} \end{cases}$$

and Y_iD is the i^{th} test result. The area A is equivalent to the **Mann-Whitney U statistic**. The proof of this is presented in Pepe [47]. In addition, the discrete nature of the empirical ROC curve causes interpretation of the variance of $\hat{A}$ to be complicated. The analytic expression for the asymptotic variance is given in Pepe [47] by

$$\text{Var}(\hat{A}) = \frac{A(1-A) + (n_0 - 1)(Q_1 - A^2) + (n_1 - 1)(Q_2 - A^2)}{n_0 n_1}$$

where

$$Q_1 = \frac{A}{2 - A}$$
$$Q_2 = \frac{2A^2}{1 + A}$$

The empirical ROC curve preserves many properties of that of the theoretical curve; in fact it is uniformly convergent to the theoretical curve [citation needed[Luzia Gon calves]. However, as expected, the empirical ROC curve has some drawbacks. In particular, it may suffer from large variability when using small sample sizes, as often the case in medical studies. Furthermore, the *jaggedness* only leads to the belief that the empirical approach is trying to estimate a smooth ROC curve.

Confidence Intervals for AUC values Due to the discrete nature of $\hat{R}(t)$, the variance of the AUC is often complicated. Analytic solutions have been well established, and the following results can be found in Pepe [47]. We wish to add 95% confidence interval for $A\hat{U}C(\hat{R}(t))$. For large samples, the area is approximately normally distributed[cite delong]. Hence a 95% confidence interval can by computed using the standard normal distribution

$$A \pm z_{\frac{\alpha}{2}} \text{SE}(A)$$

The formula for SE(A), as given by Pepe [47] and used in the popular R package `pROC`, is

$$\text{SE}(A) = \sqrt{\text{Var}\hat{A}}$$

6.5 Semiparametric Methods

The parametric approach requires the harsh assumption that the distributions of test results for both populations be Guassian. It generates a ROC curved based on the normal distribution. However this is a nuisance because we are interested in the relationship between the distribtuions of X and Y, and not with the distributions themselves. On the other extreme, the empirical method does not require any assumptions, but is inherently weaker than parametric method in terms of interpretation and analysis. In fact, the discrete ROC curve constructed from empirical data may even break certain *nice* properties of ROC curves. The semiparametric approach is such that it models the ROC curve as a smooth parametric function, rather than modelling the probability distribution. These are also known as *parametric-distribution free* methods. This approach produces a smooth ROC curve while requiring none of the harsh assumptions on the underlying test score distribution.

There are many semiparameteric methods, including Maximum Likelihood, Gaussian Mixture Models, Generalized Linear Models and Kernel Estimators. For our work, we focus on the ROC-GLM semiparametric method by [47, 11]. The GLM estimates the parameters a, b and the corresponding $\hat{A}$ [47, 11]. Consider the binary indicator variable [11]

$$U_{ij} = I(y_i, x_j), \quad i = 1 \dots n_D, j = 1 \dots n_N$$

for all $n_D \times n_N$ pairs of diagnostic test results. This indicator variable gives an alternative representation of the ROC curve Pepe [47]

$$E[U_{ij}] = R(t)$$

The ROC curve is then constructed parametrically as

$$g(R(t)) = \sum_s \beta_s h_s(t)$$

where g is the link function, h_s are the basis functions and the $beta_s$ are the unknown parameters. Note that if we use $g = \Phi^{-1}$, the probit link, $h_1(t) = 1$, and $h_2(t) = \Phi^{-1}(t)$, we retrieve the binormal method as defined in Equation 6.5. Therefore, the we have a linear model:

$$R(t) = E(U_{ij}) = \Phi(\beta_1 + \beta_2 \Phi^{-1}(t_j)) \tag{6.11}$$

where t_j are the false positive rates as shown in Colak et al. [11]. The parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are calculated using standard regression frameworks and can be used for $\hat{a}$ and $\hat{b}$. Since we forced $R(t)$ to assume a parametric form, the corresponding AUC is given by

$$\hat{A} = \Phi\left(\frac{\hat{\beta}_1}{\sqrt{1+\hat{\beta}_2}}\right)$$

6.6 Simulation Studies

Random datasets were generated from the normal distribution was used to compare and analyze the performance of parametric, nonparametric, and semiparametric methods. In particular, diagnostic test results were generated for both non-diseased (X) and diseased (Y) where $X \sim N(0,1)$ and $Y \sim N(a/b, 1/b)$, with $a = 1.4$ and $b = 0.9$. The corresponding AUC, given by Equation 6.6, is ≈ 0.850. Different ROC methods were applied to this dataset and the summary measures are recorded.

6.7 Discussion

This chapter presents a succinct introduction to the statistical modeling of ROC curves. There exists different methods to estimate ROC curves and its summary statistics. In general, these methods can be grouped into parametric, semiparametric and nonparametric forms. Parametric methods assume that the two binary populations follow a certain distribution and derives a closed form expression for the ROC curve. In other words, the distribution of the test scores completely determines the ROC curve [47]. The Gaussian family of distributions offer an obvious and simple choice. Parametric approaches are, in nature, theoretical but offer simplicity and a means of understanding the concept. The smoothness of the curve and the small number of parameters involved allow for a clear and concise exposition.

The semiparametric method for constructing ROC curves is a viable alternative to parametric and nonparametric ROC methods. The parametric method requires that the distribution of the diagnostic test be known. On the other hand, the nonparametric method may not yield a *proper* nor a smooth ROC curve, especially in small samples [47]. The semiparametric method offers an attractive approach by merging the smoothness properties of parametric methods and components from the nonparametric methods. This method requires

no assumption to be made about the distribution of the diagnostic test scores, but returns a smooth curve. GLM methods applied to the dataset reduces the problem to estimating the parameters of a Gaussian distribution, ie a, b, and AUC [11]. The use of of flexible models, such as Bayesian methods, ML estimation, and Monte Carlo simulations all offer a valid way to estimate the parameters. However, like any estimation problem, the lack of fit is a potential issue for semiparametric methods [11]. In conclusion, Pepe [47] and Colak et al. [11] show that semiparametric ROC analysis by GLM application is a reliable method that can be used as an alternative to parametric and nonparametric methods.

	Empirical					Semiparametric			
n_0/n_1	AUC	SE(AUC)	95% CI	a	b	AUC	95% CI	a	b
10	0.8	0.102	0.5893-1	-	-	0.7743	0.5026-0.9011	1.518149	-1.750568
25	0.8528	0.055	0.745 - 0.96	-	-	0.8466	0.7258-0.9353	1.617959	-1.227019
50	0.8728	0.036	0.8015-0.9441	-	-	0.8677	0.7899-0.9339	1.3269	-0.643
100	0.8856	0.024	0.8403-0.9309	-	-	0.8829	0.8343-0.9258	1.831604	-1.17063
200	0.8375	0.02	0.7993-0.8758	-	-	0.8396	0.8002-0.8748	1.3736053	-0.9563

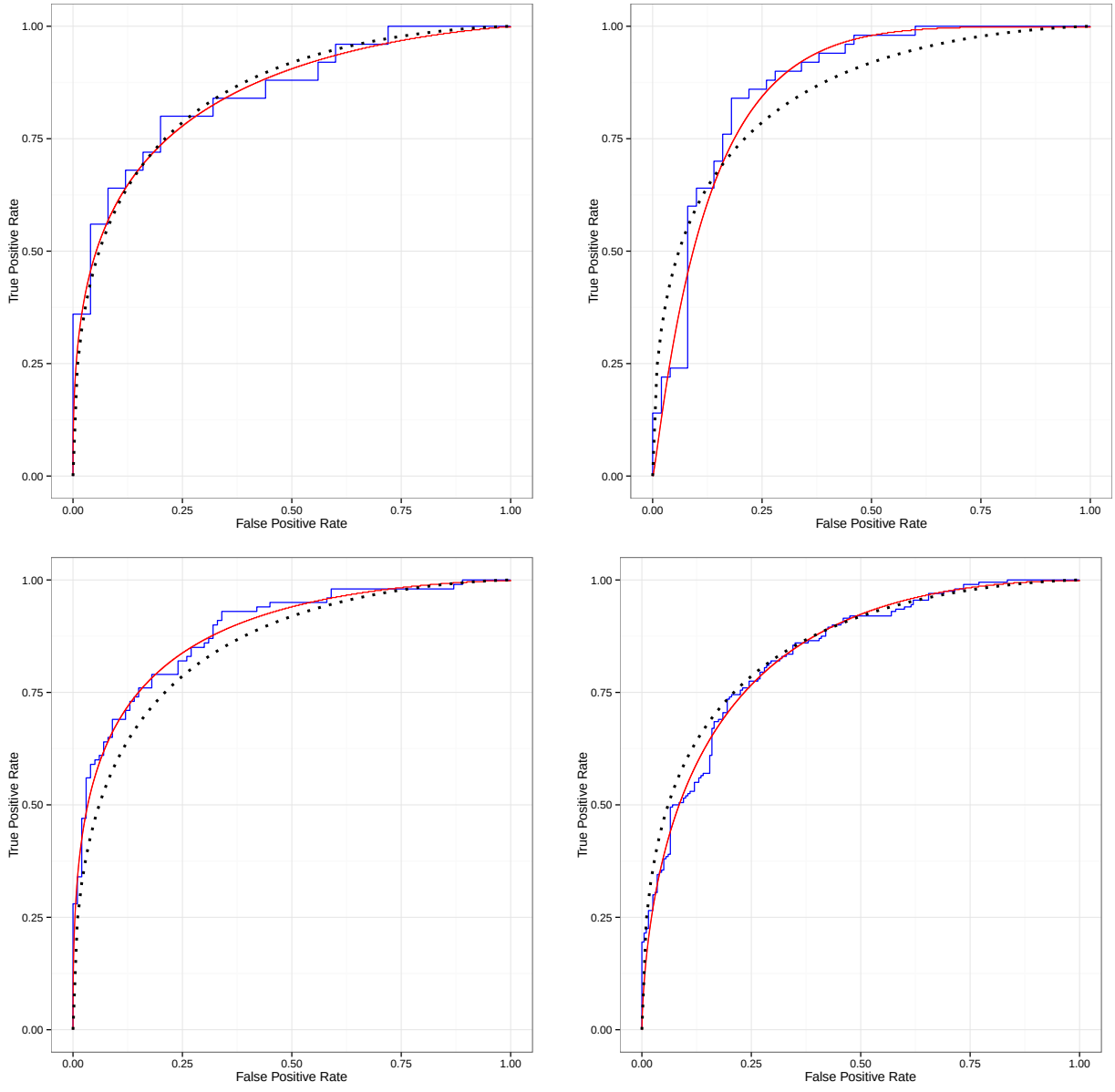


Table 6.1: Comparison between the parametric, semiparametric, and nonparametric (empirical) methods. The table shows the 95% confidence band for the AUC.

Bibliography

[1] Uri Alon. "How To Choose a Good Scientific Problem". In: *Molecular Cell* 35.6 (2009), pp. 726–728. ISSN: 10972765. arXiv: z0024.

[2] V. Andres, M. Cervera, and V. Mahdavi. "Determination of the Consensus Binding Site for MEF2 Expressed in Muscle and Brain Reveals Tissue-specific Sequence Constraints". In: *Journal of Biological Chemistry* 270.40 (Oct. 1995), pp. 23246–23249. ISSN: 0021-9258. DOI: 10.1074/jbc.270.40.23246. URL: http://www.jbc.org/content/270/40/23246.full.

[3] Christian Arnold, Peter F. Stadler, and Sonja J. Prohaska. "Chromatin computation: Epigenetic inheritance as a pattern reconstruction problem". In: *Journal of Theoretical Biology* 336 (2013), pp. 61–74. ISSN: 00225193. DOI: 10.1016/j.jtbi.2013.07.012.

[4] Patrik Asp et al. "Genome-wide remodeling of the epigenetic landscape during myogenic differentiation." In: *Proceedings of the National Academy of Sciences of the United States of America* 108 (2011), E149–E158. ISSN: 0027-8424. DOI: 10.1073/pnas.1102223108.

[5] Adrian Baddeley and Rolf Turner. "spatstat: An R Package for Analyzing Spatial Point Patterns". In: *Journal Of Statistical Software* 12.6 (2005), pp. 1–42. ISSN: 15487660. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.126.8464.

[6] Timothy L. Bailey et al. "MEME Suite: Tools for motif discovery and searching". In: *Nucleic Acids Research* 37.SUPPL. 2 (2009). ISSN: 03051048. DOI: 10.1093/nar/gkp335.

[7] B L Black and E N Olson. "Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins." In: *Annual review of cell and developmental biology* 14 (1998), pp. 167–196. ISSN: 1081-0706. DOI: 10.1146/annurev.cellbio.14.1.167.

[8] Russell Bonneville and Victor X Jin. "A hidden Markov model to identify combinatorial epigenetic regulation patterns for estrogen receptor α target genes." In: *Bioinformatics (Oxford, England)* 29

(2013), pp. 22–8. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts639. URL: http://www.ncbi.nlm.nih.gov/pubmed/23104890.

[9] David M Budden et al. "Predicting expression: the complementary power of histone modification and transcription factor binding data". In: *Epigenetics & Chromatin* 7.1 (2014), p. 36. ISSN: 1756-8935. DOI: 10.1186/1756-8935-7-36. URL: http://www.epigeneticsandchromatin.com/content/7/1/36.

[10] J M Claverie and S Audic. "The statistical significance of nucleotide position-weight matrix matches." In: *Computer applications in the biosciences : CABIOS* 12 (1996), pp. 431–439. ISSN: 0266-7061. DOI: 10.1093/bioinformatics/12.5.431.

[11] Ertugrul Colak et al. "Comparison of semiparametric, parametric, and nonparametric ROC analysis for continuous diagnostic tests using a simulation study and acute coronary syndrome data". In: *Computational and Mathematical Methods in Medicine* 2012 (2012). ISSN: 1748670X. DOI: 10.1155/2012/698320.

[12] Gabriel Cuellar-Partida et al. "Epigenetic priors for identifying active transcription factor binding sites". In: *Bioinformatics* 28 (2012), pp. 56–62. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr614.

[13] Modan K Das and Ho-Kwok Dai. "A survey of DNA motif finding algorithms." In: *BMC bioinformatics* 8 Suppl 7 (2007), S21. ISSN: 14712105. DOI: 10.1186/1471-2105-8-S7-S21.

[14] J. Davila-Velderrain, J. C. Martinez-Garcia, and E. R. Alvarez-Buylla. *Epigenetic Landscape Models: The Post-Genomic Era*. en. Tech. rep. Apr. 2014, p. 004192. DOI: 10.1101/004192. URL: http://biorxiv.org/content/early/2015/02/28/004192.abstract.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 39.1 (1977), pp. 1–38. ISSN: 0035-9246. DOI: 10.1.1.133.4884. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.133.4884.

[16] Colin Dewey. "Learning Sequence Motif Models Using Gibbs Sampling Goals for Lecture". In: (2014).

[17] Marko Djordjevic, Anirvan M. Sengupta, and Boris I. Shraiman. "A biophysical approach to transcription factor binding site discovery". In: *Genome Research* 13.11 (2003), pp. 2381–2390. ISSN: 10889051. DOI: 10.1101/gr.1271603.

[18] Annette J Dobson. *An introduction to generalized linear models*. 2002, p. 225. ISBN: 1-58488-165-8.

[19] M. Downes. "Short Math Guide for LATEX". In: *American Mathematical Society* (2002). Ed. by E Rogers and J O'Reilly, pp. 1–17. DOI: 10.1.1.96.645.

[20] S R Eddy. “Profile hidden Markov models.” In: *Bioinformatics (Oxford, England)* 14 (1998), pp. 755–763. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/14.9.755.

[21] Sean R Eddy. “What is a hidden Markov model?” In: *Nature biotechnology* 22 (2004), pp. 1315–1316. ISSN: 1087-0156. DOI: 10.1038/nbt1004-1315.

[22] Ivan Erill and Michael C O’Neill. “A reexamination of information theory-based methods for DNA-binding site identification.” In: *BMC bioinformatics* 10 (2009), p. 57. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-57.

[23] Jason Ernst and Manolis Kellis. *ChromHMM: automating chromatin-state discovery and characterization*. 2012. DOI: 10.1038/nmeth.1906.

[24] Jason Ernst and Manolis Kellis. “Discovery and characterization of chromatin states for systematic annotation of the human genome.” In: *Nature biotechnology* 28 (2010), pp. 817–825. ISSN: 1087-0156. DOI: 10.1038/nbt.1662.

[25] Jason Ernst et al. “Integrating multiple evidence sources to predict transcription factor binding in the human genome”. In: *Genome Research* 20.4 (2010), pp. 526–536. ISSN: 10889051. DOI: 10.1101/gr.096305.109.

[26] Jason Ernst et al. “Mapping and analysis of chromatin state dynamics in nine human cell types.” In: *Nature* 473 (2011), pp. 43–49. ISSN: 0028-0836. DOI: 10.1038/nature09906.

[27] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (June 2006), pp. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010. URL: http://www.sciencedirect.com/science/article/pii/S016786550500303X.

[28] J W Fickett. “Coordinate positioning of MEF2 and myogenin binding sites.” In: *Gene* 172 (1996), GC19–C32. ISSN: 03781119. DOI: 10.1016/0378-1119(95)00888-8.

[29] J W Fickett. “Quantitative discrimination of MEF2 sites.” In: *Molecular and cellular biology* 16.1 (1996), pp. 437–441.

[30] Naum I. Gershenzon, Gary D. Stormo, and Ilya P. Ioshikhes. “Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites”. In: *Nucleic Acids Research* 33.7 (2005), pp. 2290–2301. ISSN: 03051048. DOI: 10.1093/nar/gki519.

[31] Mickael Goujon et al. “A new bioinformatics analysis tools framework at EMBL-EBI”. In: *Nucleic Acids Research* 38 (2010). ISSN: 03051048. DOI: 10.1093/nar/gkq313.

[32] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. "FIMO: Scanning for occurrences of a given motif". In: *Bioinformatics* 27 (2011), pp. 1017–1018. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr064.

[33] Sridhar Hannenhalli. "Eukaryotic transcription factor binding sites–modeling and integrative search methods." In: *Bioinformatics (Oxford, England)* 24 (2008), pp. 1325–1331. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btn198.

[34] Steven Henikoff, Shmuel Pietrokovski, and Jorja G. Henikoff. "Superior performance in protein homology detection with the Blocks Database servers". In: *Nucleic Acids Research* 26.1 (1998), pp. 309–312.

[35] G. Z. Hertz and G. D. Stormo. "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences". In: *Bioinformatics* 15.7 (July 1999), pp. 563–577. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/15.7.563. URL: http://bioinformatics.oxfordjournals.org/content/15/7/563.short.

[36] J M Heumann, A S Lapedes, and G D Stormo. "Neural networks for determining protein specificity and multiple alignment of binding sites." In: *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* 2 (1994), pp. 188–194. ISSN: 1553-0833.

[37] Michael M. Hoffman et al. "Integrative annotation of chromatin elements from ENCODE data". In: *Nucleic Acids Research* 41 (2013), pp. 827–841. ISSN: 03051048. DOI: 10.1093/nar/gks1284.

[38] Dana S F Homsi, Vineet Gupta, and Gary D Stormo. "Modeling the quantitative specificity of DNA-binding proteins from example binding sites." In: *PloS one* 4.8 (Jan. 2009), e6736. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0006736. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2726951%5C&tool=pmcentrez%5C&rendertype=abstract.

[39] Hideya Kawaji et al. "Update of the FANTOM web resource: From mammalian transcriptional landscape to its dynamic regulation". In: *Nucleic Acids Research* 39 (2011). ISSN: 03051048. DOI: 10.1093/nar/gkq1112.

[40] A Krogh. "An introduction to hidden Markov models for biological sequences". In: *New Comprehensive Biochemistry* (1998). URL: http://www.kcl.ac.uk/ip/philcunningham/bmsr/nic12/HMM-Krogh.pdf.

[41] Jessica L. Larson et al. "A tiered hidden Markov model characterizes multi-scale chromatin states". In: *Genomics* 102 (2013), pp. 1–7. ISSN: 08887543. DOI: 10.1016/j.ygeno.2013.03.009.

[42] Leping Li, Yu Liang, and Robert L Bass. "GAPWM: a genetic algorithm method for optimizing a position weight matrix." In: *Bioinformatics (Oxford, England)* 23.10 (May 2007), pp. 1188–94. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm080. URL: http://www.ncbi.nlm.nih.gov/pubmed/17341493.

[43] Henrik Madsen and Poul Thyregod. *Introduction to General and Generalized Linear Models*. 2011.

[44] Robert C. McLeay et al. "Genome-wide in silico prediction of gene expression". In: *Bioinformatics* 28 (2012), pp. 2789–2796. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts529.

[45] C E Metz. "Basic principles of ROC analysis." In: *Seminars in nuclear medicine* 8.4 (Oct. 1978), pp. 283–98. ISSN: 0001-2998. URL: http://www.ncbi.nlm.nih.gov/pubmed/112681.

[46] Keishin Nishida, Martin C. Frith, and Kenta Nakai. "Pseudocounts for transcription factor binding sites". In: *Nucleic Acids Research* 37 (2009), pp. 939–944. ISSN: 03051048. DOI: 10.1093/nar/gkn1019.

[47] Margaret Sullivan Pepe. *Statistical Evaluation of Medical Tests and Biomarkers for Classification*. 2007.

[48] Lawrence R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77 (1989), pp. 257–286. ISSN: 00189219. DOI: 10.1109/5.18626. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=18626.

[49] Benjamin Schuster-Böckler and Alex Bateman. "An introduction to hidden Markov models." In: *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Appendix 3 (2007), Appendix 3A. ISSN: 1934-340X. DOI: 10.1002/0471250953.bia03as18.

[50] Fabian Sievers et al. *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega*. 2011. DOI: 10.1038/msb.2011.75.

[51] Johannes Söding. "Protein homology detection by HMM-HMM comparison". In: *Bioinformatics* 21 (2005), pp. 951–960. ISSN: 13674803. DOI: 10.1093/bioinformatics/bti125.

[52] R Staden. "Methods for calculating the probabilities of finding patterns in sequences." In: *Computer applications in the biosciences : CABIOS* 5.2 (1989), pp. 89–96. ISSN: 0266-7061. DOI: 10.1093/bioinformatics/5.2.89.

[53] John D Storey and Robert Tibshirani. "Statistical significance for genomewide studies." In: *Proceedings of the National Academy of Sciences of the United States of America* 100 (2003), pp. 9440–9445. ISSN: 0027-8424. DOI: 10.1073/pnas.1530509100.

[54] G D Stormo. "Consensus patterns in DNA." In: *Methods in enzymology* 183 (1990), pp. 211–221. ISSN: 0076-6879.

[55] G D Stormo. "DNA binding sites: representation and discovery." In: *Bioinformatics (Oxford, England)* 16 (2000), pp. 16–23. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/16.1.16.

[56] G D Stormo and D S Fields. "Specificity, free energy and information content in protein-DNA interactions." In: *Trends in biochemical sciences* 23.3 (1998), pp. 109–113. ISSN: 0968-0004. DOI: 10.1016/S0968-0004(98)01187-6.

[57] Mohammad Talebzadeh and Fatemeh Zare-Mirakabad. "Transcription factor binding sites prediction based on modified nucleosomes". In: *PLoS ONE* 9 (2014). ISSN: 19326203. DOI: 10.1371/journal.pone.0089226.

[58] Hélène Touzet and Jean-Stéphane Varré. "Efficient and accurate P-value computation for Position Weight Matrices." In: *Algorithms for molecular biology : AMB* 2 (2007), p. 15. ISSN: 1748-7188. DOI: 10.1186/1748-7188-2-15.

[59] Hélène Touzet and Jean-Stéphane Varré. "Efficient and accurate P-value computation for Position Weight Matrices." In: *Algorithms for molecular biology : AMB* 2 (2007), p. 15. ISSN: 1748-7188. DOI: 10.1186/1748-7188-2-15.

[60] Anton Valouev et al. "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data". In: *Nat Methods* 5 (2008), pp. 829–834. ISSN: 1548-7091. DOI: 10.1038/nmeth.1246.

[61] C H Waddington. "The epigenotype. 1942." In: *International journal of epidemiology* 41.1 (Feb. 2012), pp. 10–3. ISSN: 1464-3685. DOI: 10.1093/ije/dyr184. URL: http://www.ncbi.nlm.nih.gov/pubmed/22186258.

[62] Stephanie Wales et al. "Global MEF2 target gene analysis in cardiac and skeletal muscle reveals novel regulation of DUSP6 by p38MAPK-MEF2 signaling." In: *Nucleic acids research* 42.18 (Oct. 2014), pp. 11349–62. ISSN: 1362-4962. DOI: 10.1093/nar/gku813. URL: http://www.ncbi.nlm.nih.gov/pubmed/25217591.

[63] Wyeth W Wasserman and Albin Sandelin. "Applied bioinformatics for the identification of regulatory elements." In: *Nature reviews. Genetics* 5 (2004), pp. 276–287. ISSN: 1471-0056. DOI: 10.1038/nrg1315.

[64] George M. Whitesides. "Whitesides' Group: Writing a paper". In: *Advanced Materials* 16.15 SPEC. ISS. (2004), pp. 1375–1377. arXiv: 1003.3921v1.

[65] Troy W Whitfield et al. "Functional analysis of transcription factor binding sites in human promoters." In: *Genome biology* 13 (2012), R50. ISSN: 1465-6914. DOI: 10.1186/gb-2012-13-9-r50. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3491394%5C&tool=pmcentrez%5C&rendertype=abstract.

[66] Xuhua Xia. "Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction". In: *Scientifica* 2012 (2012), pp. 1–15. ISSN: 2090-908X. DOI: 10.6064/2012/917540. URL: http://www.hindawi.com/journals/scientifica/2012/917540/.

[67] Haipeng Xing et al. "Genome-wide localization of protein-DNA binding and histone modification by a bayesian change-point method with ChIP-seq data". In: *PLoS Computational Biology* 8.7 (2012). ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1002613.

[68] Hong-Mei Zhang et al. "AnimalTFDB: a comprehensive animal transcription factor database." In: *Nucleic acids research* 40.Database issue (Jan. 2012), pp. D144–9. ISSN: 1362-4962. DOI: 10.1093/nar/gkr965. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245155%5C&tool=pmcentrez%5C&rendertype=abstract.

[69] Yue Zhao, David Granas, and Gary D. Stormo. "Inferring binding energies from selected binding sites". In: *PLoS Computational Biology* 5.12 (2009). ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000590.