A SEQUENCE BASED MODEL FOR NUCLEOSOME POSITIONING IN YEAST

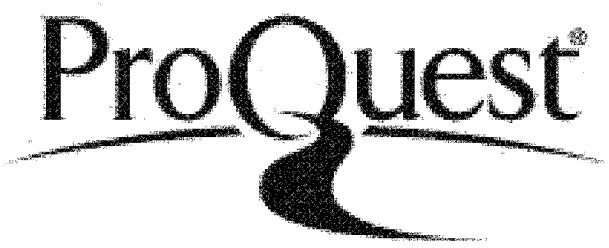
by

Mahmoud Ghandi

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

June, 2012



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Gene regulation is a central process in every living organism. At the transcriptional level, gene regulation usually involves interactions of regulatory proteins, called transcription factors (TFs), with DNA through sequence specific binding, and interactions with other TFs and the basal transcription machinery. In eukaryotic cells, nucleosome positioning presents another level of complexity to gene regulation. Nucleosomes are formed by wrapping stretches of about 147bp of DNA around a core consisting of eight histone proteins. Nucleosomes can affect gene regulation by impeding or in some cases catalyzing the interactions between transcription factors and their binding sites. Because of their prominent role in transcriptional regulation, a comprehensive description of gene expression must include a predictive model of nucleosome positioning. This dissertation focuses on developing such a model and is organized in three parts: First we present Group Normaliation, a novel microarray data normalization method that removes probe effects by using a set of reference probes with similar hybridization properties. This approach more accurately specifies nucleosome positions and changes in nucleosome occupancy compared to conventional

methods. Our results show a 1.7(dB) improvement in signal quality at probe level compared to MAS5.0 algorithm. In the second part, we present a novel model to more accurately predict nucleosome positioning in yeast from DNA sequence. This model predicts favorable nucleosome positions by a) estimating minimum energy twist and bend angles for a curved DNA molecule, and b) by indirectly modeling the DNA-histone core interactions. These two aspects are unified in a context based sequence model and the nucleosome positioning data obtained in part 1 is used to derive the model parameters. Finally in part three, we present experimental validation results for the nucleosome positioning model. We used the nucleosome positioning model to design mutations that change the nucleosome affinity at the promoter region of the yeast hexose transporter gene HXT3 and then measured its effect on HXT3 response to enviroumental glucose concentration by flouroscence microscopy and compared the results with the model predictions.

Advisor: Michael A. Beer, PhD

Thesis Readers: Michael A. Beer, PhD, Joel S. Bader, PhD

Thesis committee: Michael A. Beer, PhD, Joel S. Bader, PhD, Rachel Karchin, PhD, Feilim Mac Gabhann, PhD, Jeffry L Corden, PhD and Sarah J. Wheelan, MD, PhD

Acknowledgments

Above all I thank God the creator for endowing me the ability to understand some of the beauties of nature and for the kind and supporting family, friends and professors.

I would like to acknowledge my advisor Mike Beer for his guidance and help throughout this work, without which this dissertation would not be possible. Thank you for your mentorship and support and for allowing me to think and develop independently.

I would also like to thank members of the thesis committee, Drs. Bader, Karchin, Mac Gabhann, Corden and Wheelan for their time and for the fruitful discussions and comments.

I also want to thank everyone else whom I have interacted with including my labmates Jun Kyu Rhee, Rahul Karnik, Dongwon Lee, Donavan Cheng, Cecilia Ng, Navneeta Bansal, Yan Qi at Beer lab and Ashish Kapoor, Dallas Miller and Maria X Sosa of Chakravarti lab, Ed Davis and Boeke lab for sharing the -URA strain, Brendan Cormack for yeast optimized GFP vector, Jameson Ribbens and Maegawa

Lab for sharing the fluorescence plate reader and all my professors at Johns Hopkins School of Medicine and School of Public Health, in particular: Reza Shadmehr, Jon Lorsch, Jeremy Nathans, Rafael Irizarry, Peter Maloney, Joel Bader and Jeff Corden. I want to specifically acknowledge Morteza Mohammad-Noori for his help with the proof of the Equation 5.10 and Dongwon Lee for providing CTCF and P300 datasets and running SVM algorithm with k-mers kernel for these datasets.

And last but not least I want to thank my wife, Maryam, for her patience and unconditional support and love.

Dedication

This thesis is dedicated to my mother Simin Nekoofar, my father Mahmoud Ghandi, my wife Maryam, and our children Yasaman and Yusuf.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Cells employ many mechanisms to tightly regulate the expression of genes. This regulation is crucial for survival and optimal function. Disruption of regulatory mechanisms causes uncontrolled expression and often, severe outcomes for the organism. Some human diseases, such as cancer are associated with disordered gene expression. At the transcriptional level, gene regulation usually involves interactions of regulatory proteins, called transcription factors (TFs), with DNA through sequence specific binding, and interactions with other TFs and the basal transcription machinery. To model these interactions, our lab has developed many computational frameworks for systematically inferring combinatorial transcriptional regulatory logic and predicting regulatory elements by detecting overrepresented sequences in promoters of co-

regulated genes [1]. In eukaryotic cells, nucleosome positioning presents another level of complexity to gene regulation. Nucleosomes are formed by wrapping stretches of about 147bp of DNA around a core consisting of eight histone proteins. Nucleosomes can affect gene regulation by impeding or in some cases catalyzing the interactions between transcription factors and their binding sites; therefore directly affecting the accessibility and efficacy of a transcription factor binding site. Nucleosome positions are not static and can be modified, correlated with up- or down-regulation of genes. This makes nucleosome positioning an indispensable component of transcriptional regulatory logic in eukaryotes. Because of their prominent role in transcriptional regulation, a comprehensive description of gene expression must include a predictive model of nucleosome positioning, and developing such a model is the primary goal of this research. Microarrays and sequencing technologies have produced high resolution maps of nucleosome positions under varying experimental conditions. While this data is extremely reproducible, variations in probe hybridization efficiency can make precise inference of nuclesome occupancy difficult. In addition, the mechanisms that specify nucleosome positions are still not clear. Our current understanding of this process is that nucleosome occupancy is determined by a balance between the energetic contributions of 1) phase dependent sequence specific interactions between DNA and the histones, and 2) a sequence specific tendency to adopt the strong deformation required to wrap DNA around the nucleosome. In addition to these, there are ATP-dependent remodeling mechanisms that influence nucleosome positions *in vivo*.

In particular we will focus on nucleosome positioning in yeast, as a model organism. Yeast has a relatively compact genome (~12 million bp) and extensive nucleosome positioning studies have been done in yeast. Histones, which are the building blocks of the nucleosomes, are highly conserved proteins and understanding nucleosome positioning in yeast is likely to be at least somewhat relevant in higher organisms. To improve our understanding of nucleosome positioning and our ability to accurately predict nucleosome positions, this thesis has focused on a combined computational and experimental approach to advance many aspects of this problem.

1.2 Thesis Organization

This thesis is organized into the following parts: In chapter 2 we give background information about what nucleosomes are, why a model for nucleosome positioning is important, and how the nuclesome positioning is experimentally determined. Our ultimate goal in this work has been to improve our models of nucleosome positioning to make more accurate predictions of their affect on gene expression. As will become clear, a key component of nucleosome positioning is sequence specific direct DNA/histones interactions. Since our sequence based models require genomic data for training, we first found it necessary to develop a new microarray data normalization procedure to improve the quality of the training data. In chapter 3 we present this genomic data normalization method to improve nucleosome positioning tiling array data

and to more accurately identify those genomic regions which are differentially occupied by nucleosomes in response to environmental or developmental stimuli. Next, in chapter 4, we use the nucleosome positioning data to develop a context-based model to more accurately predict nucleosome positioning in yeast from DNA sequence. This model makes significant improvements to our ability to predict nucleosome affinity. At the same time, we were interested in how other classes of sequence based models would compare to this approach. Models using longer sequence features have the potential to be more accurate, but suffer from overfitting small training data sets. To circumvent this fundamental problem, in chapter 5 we develop a robust method for k-mer frequency estimation using gapped k-mers, and we also develop a novel sequence similarity score based on it and show some of the results of this method as applied to enhancer prediction and nucleosome affinity in chapter 6. In chapter 7 we investigate the regular average nucleosome positioning near transcription start sites in yeast and its relation to some transcription factor binding sites. Finally, in chapter 8 we describe an experimental system we developed to validate these nucleosome positioning models and their role in modulating transcriptional response in *S. cerevisiae*. In chapter 9 we give a summary of the results and a discussion of future work.

Chapter 2

Background

2.1 Nucleosome Positioning and Gene Regulation

Cells employ many mechanisms to tightly regulate expression of genes. This regulation is crucial for survival and optimal function. Disruption of regulatory mechanisms causes uncontrolled expression and often severe outcomes for the organism. Some human diseases, such as cancer are associated with disordered gene expression.

At the transcriptional level, gene regulation usually involves interactions of regulatory proteins, called transcription factors (TFs), with DNA through sequence specific binding, and interactions with other TFs and the basal transcription machinery. To model these interactions, our lab has developed many computational frameworks

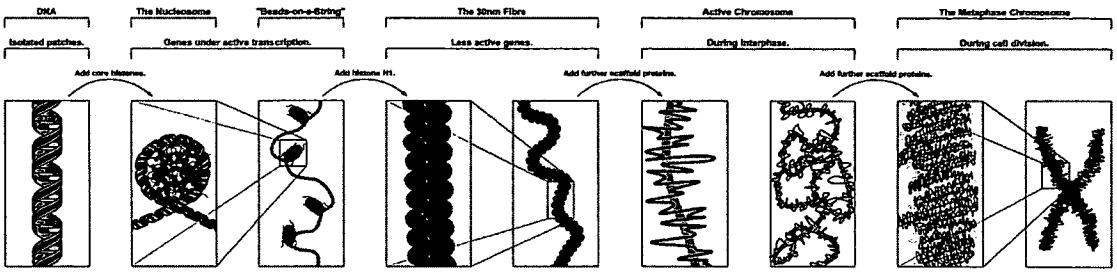


Figure 2.1: Hierarchical DNA packaging in chromatin. Nucleosomes are the first level of the chromatin organization and consist of a stretch of 147bp of DNA wrapped around a histone octamer core

for systematically inferring combinatorial transcriptional regulatory logic and predicting regulatory elements by detecting overrepresented sequences in promoters of co-regulated genes [1].

In eukaryotic cells, nucleosome positioning presents another level of complexity to gene regulation. In these cells, DNA is compacted and packaged in the nucleus in a structure consisting of a complex of DNA, RNA and protein, called chromatin [2]. As schematically depicted in Figure 2.1, nucleosomes are the building block of this structure. A nucleosome consists of a stretch of ~147 base pairs of DNA wrapped ~1.65 times around a core protein octamer [3]. The core contains two of each histone protein (H2A,H2B,H3, and H4) which are in the form of a $(H3-H4)_2$ tetramer and two H2A-H2B dimers [4, 5]. About 75%-90% of the genomic DNA is found to be nucleosome bound, the remaining consists of the nucleosome free regions that are placed between nucleosomes and are called linker DNA.

Nucleosomes are believed to affect gene regulation mostly by impeding the interac-

tion between transcription factors and their binding sites, therefore directly affecting the accessibility and efficacy of a transcription factor binding site (Figure 2.2A). Nucleosome positions are not static and can be modified, resulting in up- or down-regulation of genes (Figure 2.2B). This may occur in healthy cells in response to a stimulus or in cancerous cells as a result of unwanted genetic alterations. As an example, in the PHO regulon in S. cerevisiae, chromatin remodeling during phosphate starvation exposes binding sites for the Pho4 activator that are initially occluded by nucleosomes [6]. Also, some studies have related the epigenetically silencing of a tumor-suppressor gene in cancer cells to the presence of nucleosomes in a region in the promoter of that gene that is nucleosome free in normal cells [7]. This underscores the significance of better understanding of nucleosome positioning as a central piece in the gene regulation puzzle.

2.2 Experimental Measurement of Nucleosome Positioning

To better understand and model nucoleosome positioning, having high quality nucleosome positioning data is crucial. Recent advances in high-throughput technologies such as massively parallel sequencing and microarrays have allowed genome scale measurement of nucleosome positioning in yeast [10–14]. Similar techniques have been used to measure nucleosome positioning in *D. melanogaster* [15], *C. elegans* [16],

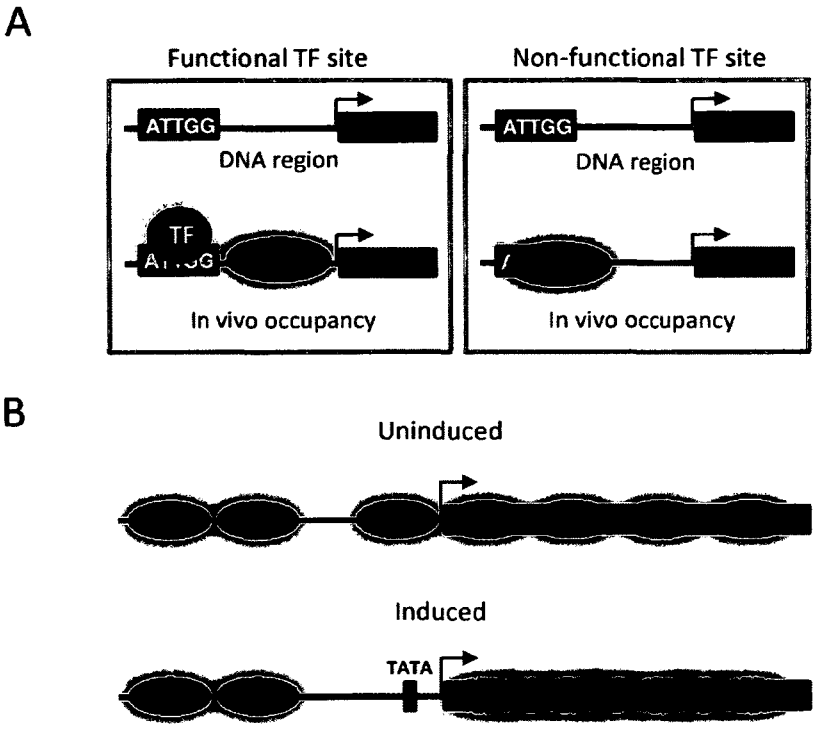


Figure 2.2: Nucleosomes can modulate TF binding. (A) On the left is a binding site that is functional and can interact with the TF. On the right, the same binding site is occupied by a nucleosome and therefore not accessible to the transcription factor. [8](B) Nucleosomes can be dynamically positioned. On the top, the TATA box of a gene is bound by a nucleosome at the uninduced condition and the gene is off. Upon transcriptional activation, the nucleosome at the promoter is removed and the gene is upregulated. [9]

Oryzias latipes (Japanese killifish) [17], and in human cells [18–20]. Figure 2.3 depicts a method commonly used to experimentally measure nucleosome positions. In this method, chromatin is cross-linked by formaldehyde and then digested by micrococcal nuclease (MNase). MNase preferentially digests regions that are not protected by a nucleosome. Then the mono-nucleosome sized (~147bp) DNA fragments are purified (usually by gel electrophoresis) and applied to microarray or high throughput sequencing to detect protected (nucleosome bound) regions. Although the sites cut by MNase have some sequence biases towards A-T rich regions [21], an MNase-independent method [22] has also generated similar maps for nucleosome positioning.

2.3 Models for Nucleosome Positioning

Nucleosome positions are not random. In particular, there is a nucleosome free region (NFR) at most of the promoters in yeast. Moreover, taking the average of nucleosome occupancy near transcription start sites (TSS) shows a significant regular nucleosome positioning pattern (Figure 2.4). The mechanism that specifies this regular nucleosome organization is still unclear.

There are different mechanisms suggested in the literature that affect nucleosome positioning. Among these are: intrinsic DNA/histone interactions [11, 24], ATP-dependent remodeling complexes [25–30], transcription factor binding and their competition with nucleosomes for DNA [31–34], and the stabilizaton or destabilization of

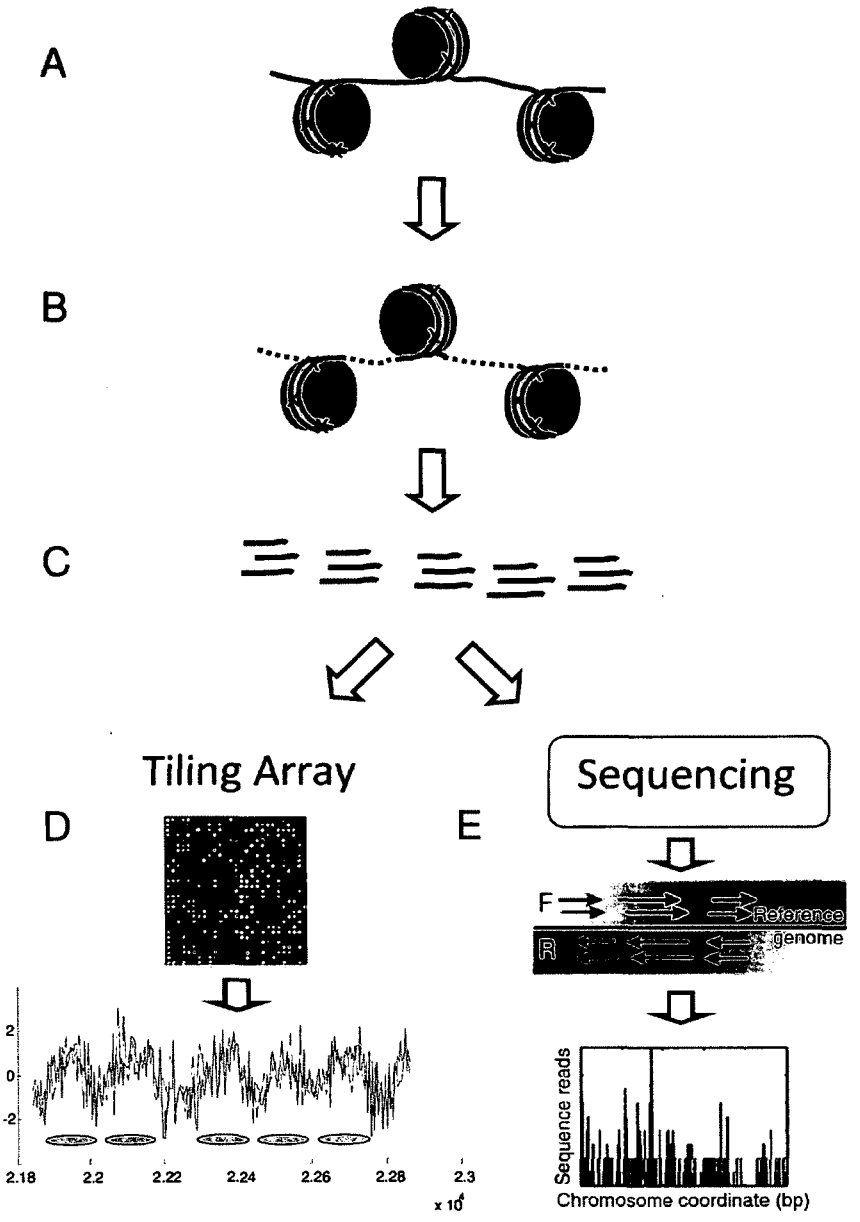


Figure 2.3: Measuring global nucleosome positioning: A) chromatin is cross-linked by formaldehyde. B) MNase preferentially digests linker DNA. C) mono-nucleosome sized (~147bp) DNA fragments are purified. D) Purified nucleosomal DNA is hybridized to tiling array or alternatively E) detected by high-throughput sequencing. Nucleosome positions are estimated by analysis of the tiling array or sequencing data [23].

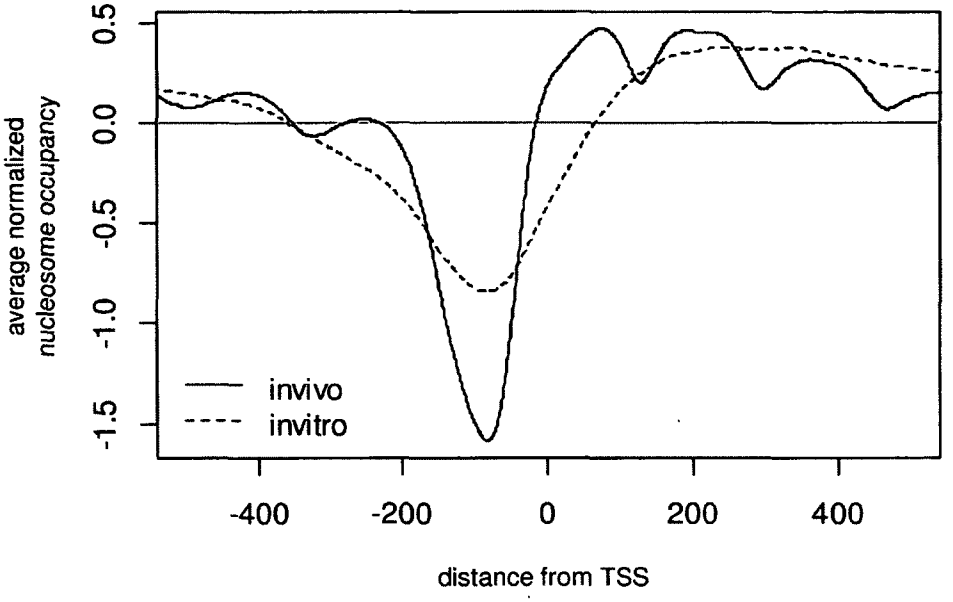


Figure 2.4: Average nucleosome occupancy profile near transcription start site in yeast: solid lines represent *in vivo* and dashed lines represent *in vitro* nucleosome occupancy profiles.

nucleosomes by combinations of these factors and RNA-PolII [35, 36].

In this section we will briefly review some of the models proposed in the literature for nucleosome positioning and then describe our approach in this thesis.

2.3.1 Intrinsic DNA/Histone Interactions

In this first class of models, intrinsic DNA-histone interactions are the dominant determinant of nucleosome positioning. In [8], Segal and his colleagues proposed that a sequence based model based on dinucleotide frequencies (obtained from experimentally specified nucleosome bound DNA) could explain up to 50% of *in vivo* nucleosome positioning in yeast. Hence, they suggested that nucleosome positions are encoded in DNA sequence as a *genomic code* and that DNA sequence and knowledge of the sequence based positioning code would be sufficient to predict nucleosome binding positions *in vivo*. These models rely on the fact that different DNA sequences have different affinities to bind to histones. This model is motivated by the observation that there is a ~10bp periodicity in the dinucleotide frequencies (in particular AA,TT,TA, and GC, where GC is placed out of phase with the other three) in nucleosome bound DNA.

Nucleosomes can also be formed *in vitro* by salt gradient dialysis [37] of purified DNA and four histone proteins. In [11], Kaplan and his colleagues used massively parallel sequencing to measure yeast genome-wide nucleosome positioning *in vivo* and *in vitro* and showed that *in vitro* nucleosome positions are very similar to *in*

vivo positions, supporting the idea that the nucleosome positions are *DNA-encoded.* In particular, the *in vitro* data also recapitulated the average nucleosome free region (NFR) upstream of transcription start sites (TSS) and near transcription termination sites (TTS). This is suggestive, but it is worth noting that the regular average nucleosome occupancy pattern near TSS (shown in Figure 2.4) was not exactly reproduced by the *in vitro* experiment, indicating that nucleosome positions are affected to some extent by other mechanisms. The 10bp periodicity of AA/TT/AT dinucleotides was observed in nucleosome sequences assembled both *in vitro* and *in vivo* [11, 14]. However, this effect was stronger *in vitro* suggesting reduced effect of rotational positioning *in vivo* [14].

In contrast to Segal's model, Kaplan's model for nucleosome positioning uses 5-mer frequencies whereas Segal's model relies only on dinucleotide frequencies. Moreover, Kaplan's model also includes contribution from sequences that are disfavoured by the nucleosomes while Segal's model was based only on nucleosomal sequences.

Another sequence based model proposed for nucleosome occupancy used wavelet transform to extract features from nucleosome bound and nucleosome free DNA sequences and built a logistic classifier to discriminate nucleosome bound regions from nucleosome free regions [38]. They used wavelet transform to capture the periodicity of different sequence features in a nucleosome, however, since they used a discrete Haar wavelet transform, this model only captures periodicity of powers of two (2,4,8,16,32,64,128) bp, while the experimentally observed features have a periodicity

of around 10bp. Also they only used the wavelet energy at each scale for each dinucleotide, and did not use the position information of individual wavelet coefficients. Similar to Segal's model, their model only relies on dinucleotide frequencies, however it also includes contribution from nucleosome free sequence features. For these reasons we believe that this approach is not the best way to model the periodicity of nucleosome sequence affinity.

Another sequence based method that was proposed for nucleosome positioning is based on support vector machines (SVM) [39]. In this model, for each sequence of length 50bp, they form a count vector that contains the counts for every k-mers for k=1,2,3,4,5,6. Then they train an SVM to classify k-mer count vectors for nucleosome bound and nucleosome free sequences. By using a k-mer frequency count vector their model loses the positional information for the k-mers. However their results show that using k-mer with $k \leq 6$, this model can discriminate bound and free sequences with high precision. Limited training sequence information and overfitting limit the maximum k they could use to build the SVM to 6. In chapter 5 we will develop a method to robustly estimate k-mer frequencies for longer k-mers using gapped k-mer frequencies and based on that in chapter 6 we will develop a sequence similarity score that we will use as a SVM kernel to model nucleosome positioning.

2.3.2 Statistical Positioning

An alternative model for nucleosome positioning that describes the regular nucleosome patterns around transcription start sites (TSS) is a model based on statistical positioning of nucleosomes near some boundary region [23, 40, 41]. The boundary region is defined by a stably positioned nucleosome or a stable nucleosome free region, and can be formed, for example, by a sequence-specific protein binding, RNA polymerase, or even a strongly nucleosome repelling sequence such as poly-T stretches. Then, although the individual nucleosomes can stochastically occupy any region of DNA (except the boundary) with similar affinity, the existence of the boundary results on average in an array of regularly spaced nucleosomes at nonrandom locations, where the nucleosomes near the boundary are well positioned but as the distance to the boundary increases, the nucleosome positions become more disordered. This is in fact somewhat consistent with the observation that *in vivo* nucleosome positions follow a regular pattern near the TSS but the regular pattern is lost after moving a distance spanning a few nucleosomes into the coding region. [42] suggested that the first well positioned nucleosome (downstream of the NFR) plays the role of a barrier and the rest of the nucleosomes are statistically positioned against this barrier. The positioning of the first nucleosome can be established by genomic sequence or by the act of some active nucleosome positioning complexes. In most of the yeast genes, the first nucleosome covers the TSS, such that the TSS is on average half a helical turn inside the first nucleosome. Moreover, it has a rotational phasing that tends to place

transcription factor binding sides on the outside surface of the nucleosome [43].

2.3.3 Nucleosomes and Transcription

The regular average pattern near the TSS is asymmetric and is directional. Also the position of the first nucleosome is closely related to transcription start site in most of the genes in yeast. For these reasons, and because of the fact that the regular nucleosome occupancy around the TSS does not appear in *in vitro* data, [14] argued against the intrinsic DNA/histone interactions as being the major determinant of nucleosome positions *in vivo*. They suggested instead that early steps in the trancription process should be the determinant of the position of the first nucleosome [14,44]. The connection between nucleosome positioning around the TSS and transcription has been the subject of several studies. It is known that the nucleosomes flanking the nucleosome free region (NFR) are enriched in H2A.Z variant [45]. H2A.Z nucleosomes are reportedly less stable *in vitro* [46], and in response to transcriptional activation, the promoter nucleosomes (including H2A.Z containing nucleosomes) may get removed [12, 18, 47, 48]. However, the fractional occupancy of H2A.Z in NFR-flanking nucleosomes is not correlated with transcription levels in yeast [49] and nucleosome-free regions in general (whether or not in promoter region) are shown to be correlated with H2A.Z.

2.3.4 ATP-dependent Factors

By using the energy from ATP hydrolysis, nucleosome remodeling complexes can also actively affect nucleosome positioning in yeast and other organisms [28, 50]. SWI/SNF complexes for example can destabilize nucleosomes in an ATP dependent manner [51]. ISWI complexes have been shown to actively slide the nucleosomes along DNA [52]. It has been shown that in ISW2 deleted yeast, nucleosomes are shifted to their favored DNA-directed position [25] and the role of Isw2 is to position nucleosomes onto unfavorable sequences. Chromatin structure remodeling (RSC) complex has also been shown to move nucleosomes away from their intrinsically favored positions. In RSC-depleted yeast, nucleosome-free regions (NFRs) shrink as the flanking nucleosomes move towards the *in vitro* model predicted sites [45]. At a subset of promoters, nucleosome positioning additionally requires the essential Myb family proteins Reb1 and Abf1 [45]. A more recent study has further highlighted the role of ATP-dependent factors on formation of regular nucleosome patterns around TSS by showing that very similar patterns form *in vitro* by salt gradient dialysis if in addition to purified DNA and histones, whole cell extract and ATP is added, but the patterns would not form in absence of ATP [53]. This shows that ATP is essential for regular positioning of nucleosomes around TSS, and intrinsic DNA/histones interactions are overcome by the ATP-dependent factors. Moreover, it can be argued that the absence of the *in vivo* regular pattern in the *in vitro* nucleosomes formed by salt gradient dialysis in the absence of ATP, as well as constant spacing between the

nucleosomes formed at a condition with lower concentration of histones, are against the statistical positioning model for nucleosomes [53].

2.3.5 Our Approach

In this thesis, we will focus on building improved sequence based models to predict nucleosome positioning by considering the geometry of nucleosome bound DNA and by developing novel computational approaches to model nucleosomes. Although precise positioning of nucleosomes requires ATP-dependent factors at many promoters in yeast, the high correlation between genomewide nucleosome occupancy *in vivo* and sequence based model predictions (learned on *in vitro* data) [11] makes it clear that intrinsic DNA/histone interactions are also playing a significant role.

Our current view of the nucleosome organization in yeast is that it is not a single mechanism that is responsible for the nucleosome positioning in all regions; instead, the final position of nucleosomes are determined by the sum of the effects of multiple mechanisms, in particular, (i) the passive intrinsic intractions of DNA and histones which forms an energy landscape for preferred nucleosome positions, and (ii) the act of ATP-dependent factors that can move nucleosomes to their unfavored positions by spending the energy of ATP hydrolysis. We will show that the former mechanism is primarily responsible for the regular nucleosome positioning pattern which is present in a subset of the genes in yeast.

On the experimental validation side, we will directly test our model with experi-

ments focusing on a glucose regulated gene, HXT3. The HXT3 promoter is differentially occupied by a nucleosome in response to glucose. At low environmental glucose level the promoter is nucleosome bound, but upon glucose addition, the nucleosome is removed and HXT3 is transcriptionally activated. This change in nucleosome occupancy clearly involves active mechanisms controlled by glucose signalling, and makes HXT3 a natural choice for investigating the effect of the intrinsic DNA/histone interactions and the ability of sequence based models to predict nucleosome affinity. Because at (or near) the critical glucose threshold at which there is a transition between nucleosome bound to nucleosome free state, the overall effect of the factors (including the ATP dependent remodeling mechanism) results in equal probability of nucleosome bound and nucleosome free states. Hence, if the change in the intrinsic sequence based binding energy of nucleosome is significant in comparison to the energy burden that the ATP-dependent factors can overcome, then we should be able to observe and measure the effect of mutations that change the intrinsic binding energy on the glucose dependent gene response.

2.4 Summary

There are different mechanisms that affect nucleosome positioning in yeast, in particular, intrinsic DNA/histone interactions and ATP-dependent mechanisms are the two main mechanisms proposed in the literature. Each of these mechanisms may

be affecting nucleosome positioning in some regions. In promoter of a subset of yeast genes, ATP-dependent mechanisms are shown to be responsible for the regular nucleosome pattern, while the intrinsic interactions of DNA and histones may determine the position of nucleosomes in some other regions of the genome not affected by ATP-dependent mechanisms. Sequence based models trained based on *in vitro* nucleosome positioning data have been shown to be able to predict the broad *in vivo* nucleosome occupancy with high precision although their ability to predict precise genomewide nucleosome positioning is challenged. In this thesis we will build improved models for nucleosome positioning by considering the geometry of nucleosome bound DNA and by developing novel computational approaches. We will also design an experimental setup to test and validate the predictions of the sequence based models.

Chapter 3

Group Normalization for Genomic Data

In [54], the effect of addition of glucose on genomewide nucleosome positioning in yeast is studied using tiling arrays. However, the tiling array signal suffers from high probe to probe variation and needs to be normalized before it can be used for further analysis. In this chapter we will develop a novel normalization method that can be used to improve the data. Data normalization is a crucial preliminary step in analyzing genomic datasets. The goal of normalization is to remove global variation to make readings across different experiments comparable. In addition, most genomic loci have non-uniform sensitivity to any given assay because of variation in local sequence properties. In microarray experiments, this non-uniform sensitivity is due to different DNA hybridization and cross-hybridization efficiencies, known as the

probe effect. In this chapter we introduce a new scheme, called Group Normalization (GN), to remove both global and local biases in one integrated step, whereby we determine the normalized probe signal by finding a set of reference probes with similar responses. Compared to conventional normalization methods such as quantile normalization and physically motivated probe effect models, our proposed method is more general and better fits to nucleosome positioning data since it does not require the assumption that the underlying signal distribution be identical for the treatment and control. Moreover, this method is flexible enough to correct for nonlinear and higher order probe effects. The Group Normalization algorithm is computationally efficient and easy to implement. We also describe a variant of the Group Normalization algorithm, called Cross Normalization, which efficiently amplifies biologically relevant differences between any two genomic datasets and use that to identify the regions where nucleosome remodeling occurs.

3.1 Introduction

Recent advances in high-throughput technologies such as massively parallel sequencing and microarrays have allowed contemporary biological experiments to routinely measure changes in molecular binding or transcription throughout the genome. These experiments generate large scale genomic datasets whose accurate interpretation requires a preliminary normalization step. For example, DNA-protein interac-

tions are commonly measured by quantifying the amount of isolated labeled target DNA by hybridization to complementary oligonucleotide probes on high density tiling microarrays (ChIP-chip), or by sequencing (ChIP-seq). In [10], Affymetrix tiling arrays consisting of over 2 million oligonucleotides of 25 base pairs (bp) each, have been used to evaluate the genome wide nucleosome positioning in yeast. During the analysis of this data we noticed that the tiling array signal is generally highly reproducible; however, it suffers from high probe to probe variation. This is most strikingly shown when randomly sheared genomic DNA is hybridized to a tiling microarray (as a control). In this case, one expects to see a flat, or uniform, hybridization signal, as genomic DNA and tiling probes are present in equal amounts. In this experiment, shown in Figure 3.1, however, the signal deviates significantly from the expected flat curve, but is highly reproducible. Figure 3.1 shows the probe level signal for two independent genomic DNA hybridizations from [10]. Similar results for ChIP-seq occur because of sequence dependencies of the sequencing assay, as shown in Figure 2a of [55]. For clarity of presentation, we will focus our description on the case of microarray hybridization, and we will refer to the phenomenon of sequence specific assay efficiency as the *probe effect*, which causes probes with identical input DNA concentration to display differential hybridization intensity. This behavior can partially be explained by the fact that individual probes bind their target DNA with varying hybridization efficiency (for example because of their different GC content), but non-specific binding (NSB) and cross hybridization also contribute to differences

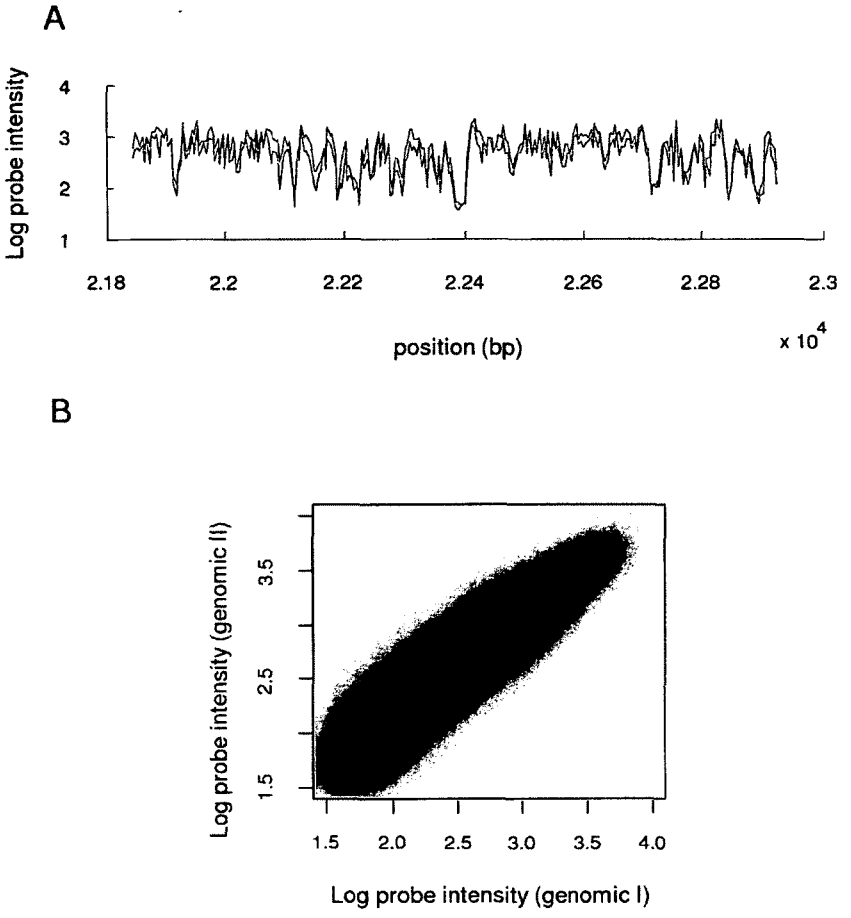


Figure 3.1: Genomic assays are often highly reproducible, but have significant efficiency variation across the genome. (A) Two genomic hybridization signals (biological replicates) from [10] shown along a portion of Chr III are highly reproducible, but deviate significantly from the expected constant signal. (B) Across the whole genome, these variations are highly reproducible. Two genomic hybridizations for the entire yeast genome are highly correlated (Pearson C=0.966).

in observed probe signals. In this chapter we propose a novel method to infer the normalized input DNA levels from microarray data and correct for these probe effects. Most other genomic data sets involve similar probe or target sequence dependencies on assay sensitivity. For example, the sensitivity of massively parallel sequencing in a ChIP-seq or RNA-seq experiment might depend on the GC content at the 5' end of the sequence. In addition to varying sequencing or hybridization efficiencies, the DNase or MNase assays used to prepare DNA may have subtle sequence affinity biases. While our method was developed by our interest in modeling nucleosome positioning data, which will be the primary example throughout the chapter, a similar approach could be applied to other genomic datasets. For example, detecting differential genomic binding of TFs due to natural variation [56] or differential gene expression from RNA-seq data [57] could benefit from our nonparametric normalization method. Improving the signal quality in genomic datasets strongly affects the accuracy and consistency of predictive models trained on this data [58].

Another consideration is that most genomic experiments involve several conditions and/or replicates. There are usually at least two conditions: treatment and control, and for each condition there may be one or multiple replicates. Different arrays might have slightly different global experimental biases (due to image scanning, variable concentrations, etc.) which are conventionally removed by a global *normalization* using one of the several available methods (see [59] for a comparison of different methods). After global normalization, the data are then *corrected for probe effect.*

Each of the normalization and probe effect correction algorithms makes assumptions about the distribution of probe signals that limits their general application. Quantile normalization [60] imposes identical probe signal distributions across conditions. This is achieved by replacing the probe signal in each condition by the mean value of probes at the same rank. Hence after quantile normalization all conditions will have the same histogram of probe signal and the Quantile-Quantile Plot (Q-Q Plot) will be a straight line. Although the assumption of identical distributions roughly holds in many cases, there are cases where it is clearly inappropriate. In the case of nucleosome positioning, about 75-80% of the genome is nucleosome bound; hence the distribution of the probe values for a nucleosome enriched condition should be significantly different than that of the genomic control.

The rank-invariant set method [61] selects a set of probes with similar rank in all conditions. This set of probes identifies regions that do not vary significantly across the different microarray conditions; for example, they may be housekeeping genes whose expression levels vary only slightly in different conditions. A nonlinear model (e.g., splines) is then fitted to the variation among the invariant set probes and used to normalize the value of all other probes. Although the assumption of invariant activity of housekeeping genes may be valid in certain cases, in many other cases it can be difficult to define a suitable invariant probe set [62, 63]. This limits the application of this normalization method.

In a more recent work, [64] have used a mixture model approach for ChIP-chip

analysis that uses LOESS curve fitting for normalization. They perform separate LOESS fitting for probes with similar GC content to better model the nonlinear relationship between probe signal on different arrays. Although this method gives promising results in a variety of ChIP-chip experiments, the strength of their method lies in its robust estimation of the null distribution. The validity of this approach is based on the assumption that the majority of probes are null, which while often the case for ChIP-chip data, is not valid for nucleosome positioning data where over 70% of the genome is nucleosome enriched. Moreover, relying only on GC content limits the flexibility of their normalization algorithm.

After performing a global condition-to-condition normalization, it is important to account for the probe effect. Earlier Affymetrix array designs attempted to correct for the probe effect by estimating non-specific binding signal (NSB), and by controlling for efficiency by choosing probes with constrained GC content. To estimate the amount of the non-specific binding signal (NSB), on some Affymetrix microarrays, there exists a mismatch probe (MM) for each perfect match probe (PM). The MM sequence is identical to that of the PM with the exception of the central base which is complementary to the central base of the PM probe. Although having a MM probe for each PM probe can help to estimate the NSB signal in some cases, in many cases MM probes does not give a direct measure of the NSB and their successful use is limited in practice. Usually, for a significant fraction of probes, the MM signal is even higher than the PM, which may be caused by the different hybridization efficiencies and

the different abundances and numbers of sequences that bind to each of the PM and MM probes. The distribution of probe signals for PM and MM probes are shown in Figure 3.2 for the [10] data, showing that they have significantly different distributions for the nucleosome enriched and genomic control conditions.

Since probe efficiency is mainly dependent on the sequence of the probes, some authors have proposed methods to directly estimate the hybridization signal (both the NSB and also the gene-specific signal (GSB)) from the probe sequence [65, 66]. In [65] the binding energies are approximated from a position-dependent weighted summation of the dinucleotide stacking energies, and least squares fitting algorithms are used to estimate the parameters. In [66] a sequence-based model for the probe affinity, called MAT, is proposed that includes a position dependent weight for each nucleotide. It also includes a nonlinear term proportional to the count of each nucleotide. It is shown that MAT can be effectively used to capture most of the probe to probe variability [66], however, even after MAT correction, the probe effect is not completely removed [67]. The reason for this is that MAT is not comprehensive enough to fully model the probe dependent effects which are known to be significantly nonlinear. In a more recent work [67] has proposed a new method called TileProbe, that employs publicly available data from the GEO database [68] to further remove the probe effect from MAT corrected intensities. TileProbe uses the median of the MAT corrected probe intensities over all samples as a model for the magnitude of the residual probe effect. Similar to [64], it relies on the assumption that most of the

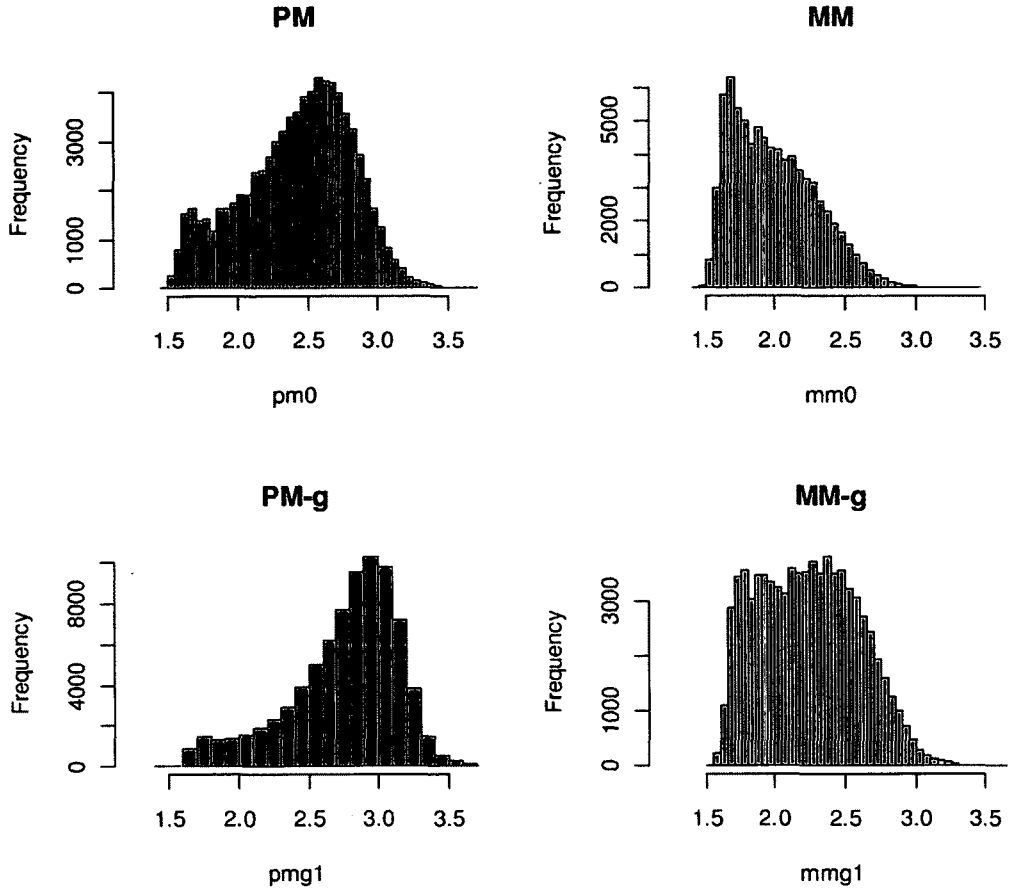


Figure 3.2: Mismatch probe distributions vary significantly in different conditions. Histograms for one treatment (nucleosome enriched, top) and one control (genomic DNA, bottom) microarray are shown. The histogram for PM (left) and MM (right) probes are plotted separately.

probe signals are generated by the null distribution. This limits the application of this method for nucleosome positioning data, where most probes have signal. Moreover, although the above mentioned sequence-based models are based on physical quantities such as stacking energies, the actual parameters are obtained by curve fitting and least square optimization, which may lead to values that are not consistent with the original model. For example in Zhang's model, in some cases, the coefficients for some dinucleotides are estimated to be positive for NSB, but negative for GSB; further evidence that the sequence-based model is overconstrained.

A comprehensive model that can explain all of the probe variations at different conditions (DNA concentrations and temperatures) would be very complicated. Here, instead of using a model-based approach, we propose a data-based approach that integrates the normalization and probe effect correction steps and eliminates the need for an explicit underlying hybridization model. Instead, we estimate the parameters of a probe's response from the response of a similar set of probes. The proposed method has the advantage of being robust and simple (no curve fitting to estimate parameters) and can effectively be used to normalize probe values. The proposed method can also be used for microarray data at two different experimental conditions to differentially amplify regions that have changed from one condition to another. We demonstrate that this method can clearly and effectively identify the biologically relevant regions in the genome which respond to an experimental stimulus, and that these regions are frequently difficult to distinguish from noise using alternative approaches.

3.2 Description of Group Normalization Method

In this section we will describe the proposed method for tiling array data normalization.

3.2.1 Group Normalization

We model the observed signal y_i for a given probe i as linear combination of three terms: first, the signal y_i is proportional to the desired biological signal, x_i, with a probe specific efficiency, A_i. Second, each probe has a background signal independent of x_i which we model as a constant signal, B_i (a combination of non-specific binding and other target independent signals), and a contribution from random noise, ε_i.

$$y_i = A_i x_i + B_i + \varepsilon_i \tag{3.1}$$

The normalized desired biological signal x_i is unitless, and can be scaled arbitrarily. In the case of nucleosome positioning, we will use $x_i = 0$ for fully unbound regions and 1 for fully nucleosome bound regions. The random noise ε_i represents all factors that cannot be modeled by A_i and B_i. The goal of normalization is to determine x_i from the observed signal y_i, and we do so by estimating A_i and B_i for each probe. Although we will focus on tiling array signals for nucleosome positioning as an example, the model given in Equation (3.1) is quite general and the normalization scheme proposed

here can be adapted to a variety of genomic assays.

As briefly discussed in Introduction, the relation between the probe effect and the probe sequence is a nonlinear and relatively complicated relation. Instead of trying to develop a physically motivated model of this relation, in our approach, Group Normalization, we model this relation implicitly from the data. This is in contrast to most model based approaches, where an explicit model is assumed, and the model parameters are estimated by fitting to the data (e.g. [65,66]). Our method relies on the fact that on each high density microarray, there exist a very large number of probes (in a typical Affymetrix oligonucleotide tiling array, there are more than 2 million probes). Figure 3.3, shows a flowchart of the proposed method. In this method, for each probe p_i, we find a set of reference probes, denoted $Ref(p_i)$, that have similar probe effects to p_i (i.e. for all the probes p_j in $Ref(p_i)$, A_j is similar to A_i and B_j is similar to B_i). The key to our Group Normalization is a ranking method to define such a reference probe set. If $Ref(p_i)$ is large enough, (we typically use $N = 1000$ probes for the reference set) despite random variation in individual probes, the probe dynamic parameters A_i and B_i can be robustly estimated from the reference set probe intensities. Figure 3.4 outlines this idea for Group Normalization for a single reference condition; below we will also consider the case of multiple reference conditions. In Figure 3.4 we highlight this process for two probes, one with high signal and one with low signal, but the procedure illustrated here is applied to all probes. First, all probes are sorted by their intensity in a reference condition, e.g., a genomic DNA reference

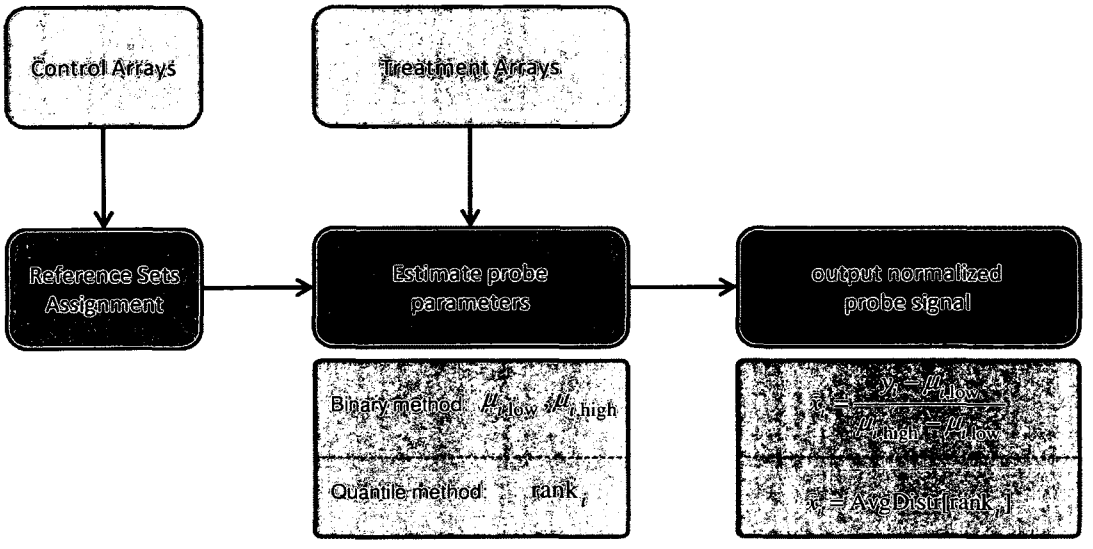


Figure 3.3: Flowchart of Group Normalization. Control arrays are used to generate reference probe sets for each probe. Then we use the reference probe sets to estimate the probe parameters in the treatment arrays and to generate the normalized signal. We propose two distinct methods to normalize the arrays: a Binary method which parameterizes high and low signal for each probe ($\mu_{i,low}$, $\mu_{i,high}$); or a Quantile-based method which uses the rank of each probe in the reference set.

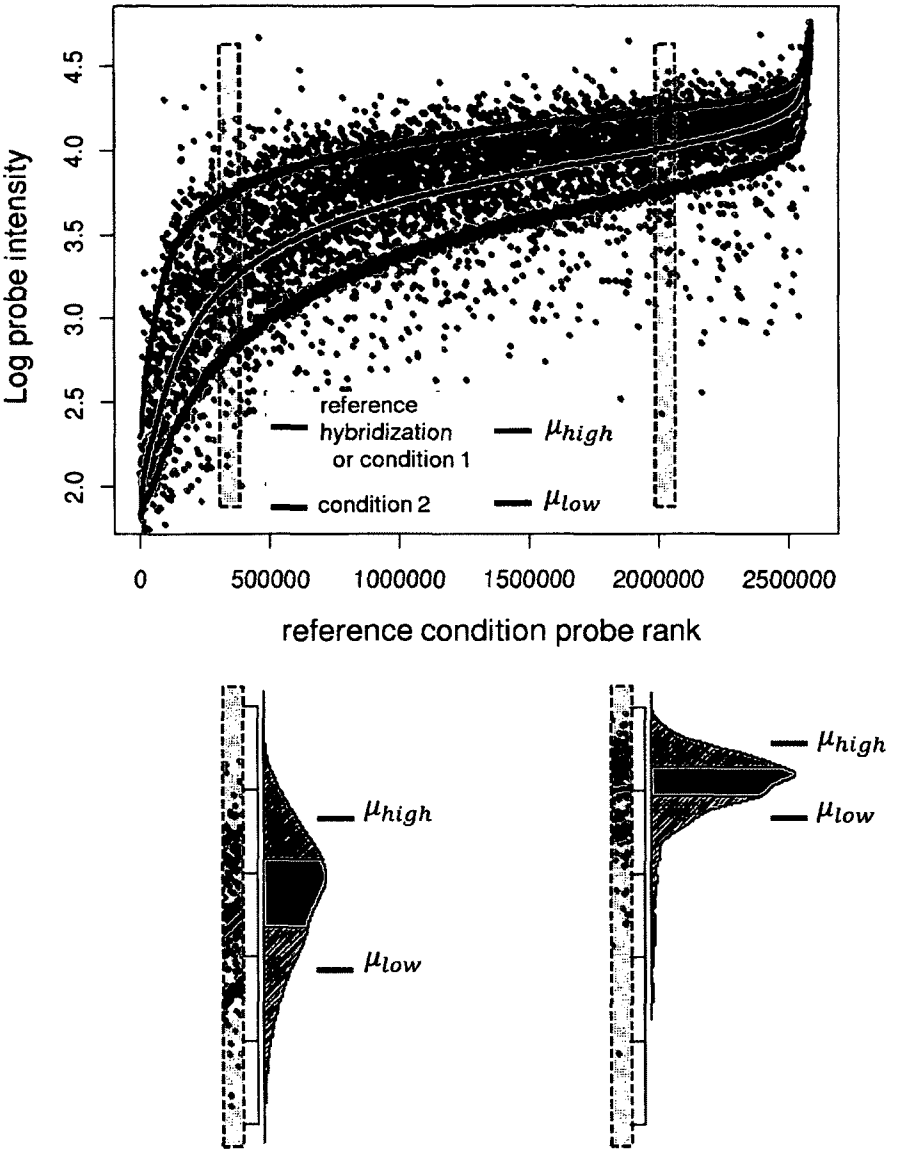


Figure 3.4: Overview of Group Normalization. Probes are shown sorted by on their values in a genomic hybridization (reference condition, black). For each probe, $N = 1000$ probes with closest signal in the genomic hybridization are assigned as reference set (dashed boxes) for each probe. Then high (red) and low (green) signal levels in the experimental condition (grey) are estimated from high and low probe signal ranges for each set of reference probes.

hybridization. For each probe i, the 1000 probes with similar rank in the reference condition define the reference set $Ref(p_i)$, shown in light blue. Then the reference set probes are sorted again by their value in a second condition, the experimental condition, e.g., nucleosome bound DNA. After sorting, the 1000 reference set probes define a mean for low signal probes, $\mu_{i,low}$, and high signal probes, $\mu_{i,high}$, within this reference set. The range of probe ranks which will define $\mu_{i,low}$ and $\mu_{i,high}$ are parameters chosen to be appropriate for the given application. In the case shown in Figure 3.4, we use the 30% lowest and highest ranking probes to define $\mu_{i,low}$ and $\mu_{i,high}$, i.e., ranks 1-300 define low signal probes and 701-1000 define high signal probes within the reference set. Finally, the normalized probe value in the experimental condition is given by:

$$x_{i,normalized} = \frac{y_i - \mu_{i,low}}{\mu_{i,high} - \mu_{i,low}} \tag{3.2}$$

Our final results are insensitive to the definition of the high and low probe ranges. The simple procedure in Equation (3.2) estimates the dynamic parameters A_i and B_i through $\mu_{i,low}$ and $\mu_{i,high}$, and explicit values for A_i and B_i are not directly required. The basic idea is that within each reference set, there are probes with high signal in the experimental condition, and there are probes with low signal in the experimental condition, and these probes effectively determine A_i and B_i for the probe p_i. We refer to the method explained above as the binary group normalization method since we estimate two signal levels (high and low) for each probe. In the following we explain an alternative approach, that does not involve estimation of $\mu_{i,low}$ and $\mu_{i,high}$.

Instead of specifying a range of probes to determine $\mu_{i,low}$ and $\mu_{i,high}$, a related approach would be to assume that the reference probes have the same distribution of the biological signal x_i. Instead of defining the ranges for low and high probes, we apply quantile normalization, and use the signal in the quantile normalized average distribution as the normalized signal. In other words, we take the average of the reference set distribution for all the probes to find an average reference set distribution, then for each probe, the rank in the reference set is calculated and the normalized signal would be the value of the average reference set distribution corresponding to that rank. We have implemented this more general approach, which we refer to as quantile-based group normalization. Compared to binary group normalization, this method gives slightly better results with spike-in ChIP-chip dataset, but we get less signal to noise improvement in nucleosome positioning data.

3.2.2 Copy Number Variation

In the definition of the reference set, we are implicitly assuming that all regions of the genome are responding to the same input DNA. But probes within repetitive regions may be responding to multiple copies of identical DNA throughout the genome, as shown in Figure 3.5 for the response of a region containing a Ty1 transposon in yeast to a genomic hybridization. Therefore, for purposes of defining the reference set probes, we explicitly ignore repeats, as these regions could reduce the accuracy of our estimation of the probe parameters.

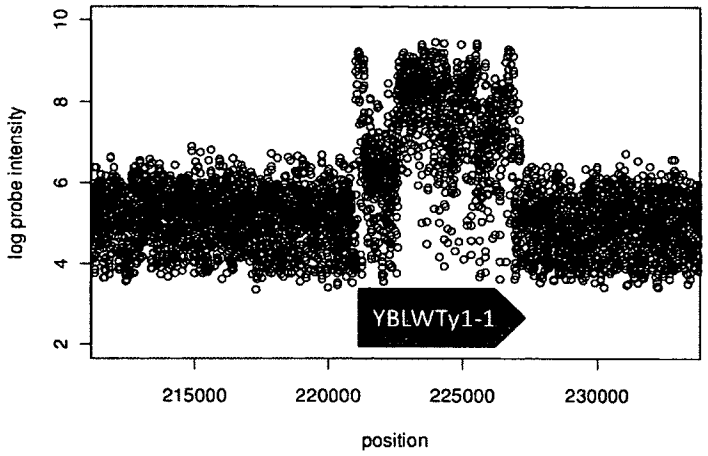


Figure 3.5: Repetitive elements have large variations in probe signal and are removed from the reference set computation. Raw probe signals for genomic hybridization (control) near YBLWTy1-1 locus on chromosome II in yeast are shown.

3.2.3 Reference Group Assignment

Group Normalization is based on finding a set of reference probes that have similar probe effects for each probe. In principle, this set could be found with high precision and provided by the chip manufacturer. But in practice, each laboratory's hybridization protocols might be applied under somewhat different conditions, and probes that have similar dynamic parameters in one lab might have different dynamic parameters in another lab. Or more generally, a good locally linear estimation of the probe parameters under some conditions might not apply to all conditions. We therefore recommend that individual users use a control reference condition, as described above, or a set of control reference conditions. Here we propose three methods to find such reference sets: 1) a single reference method, 2) a sorted average method

for multiple reference conditions, and 3) a minimum distance method for multiple reference conditions.

3.2.3.1 Single Reference Method

In this approach we perform one genomic hybridization experiment (the reference condition), as described above. In this experiment, every probe should measure an equal amount of DNA. However, since oligonucleotide probes have different affinities and dynamic properties, the measured signal is not uniform (see Figure 3.1). To define the reference set we simply sort all the probes based on their value in the reference condition and for each probe we assign the N neighboring probes as the reference set (Figure 3.4).

3.2.3.2 Sorted Average Method

Since any single hybridization is susceptible to some amount of random noise, using multiple reference hybridization may provide a more robust estimation of an appropriate reference set. When multiple reference hybridizations (conditions) are available, we propose using the sorted average method, where we sort all probes by their average signal in multiple reference hybridizations. We then assign N neighboring probes as the reference set for each probe.

3.2.3.3 Minimum Distance Method

The sorted average method minimizes the bias in estimation, but by only using the average signal, we are losing data about the variance of the probe values in the multiple reference conditions, which also gives us information about a probe's reliability and our confidence in our estimation of its parameters. To improve the reference set assignment, the minimum distance method selects N nearest probes in the multidimensional reference condition space. For example, say we have performed M genomic control hybridizations, then the distance between two probes i and j in M-dimensions is defined to be $d_{ij}^2 = (1/M)\sum_{m=1}^{M}(X_i^m - X_j^m)^2$, where $(X_i^m$ is the i'th probe signal in the m'th reference condition. So for each probe p_i, we assign the N probes with minimum distance to p_i to the reference set. This way, the probability that a highly variable probe (unstable from dataset to dataset) might mistakenly be assigned as a reference probe is decreased. Figure 3.6 depicts an example where the sorted average method and min distance methods give different results.

3.2.4 Cross Normalization

In many applications, we are interested in detecting significant differences between two conditions, neither of which is a genomic control. For example, we may be interested in comparing the early and late response to a stimulus, comparing pre- and post-stimulus, or comparing the response to two different stimuli. If a genomic

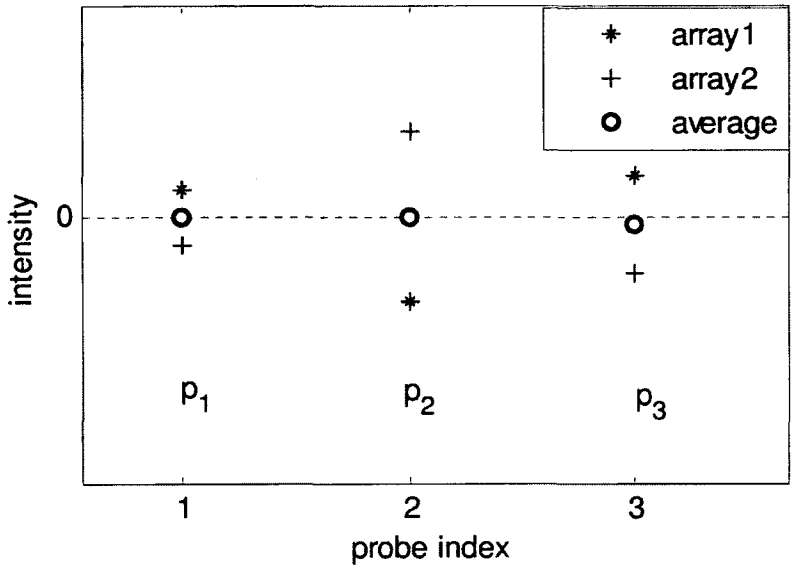


Figure 3.6: When multiple conditions are available, minimizing distance yields a more reliable reference set assignment. Here for three different probes: p1=(0.1,-0.1), p2=(-0.3,0.3), p3=(0.15,-0.20), the averages are Avg(p1) = 0; avg(p2) = 0; avg(p3) = -0.025; but distances are d12 = 0.4, d13 = 0.1. So although probe 1 and 2 have more similar averages, probe 1 and 3 have more similar responses, and probe 3 is therefore a better reference probe for probe 1.

hybridization is available, we could separately normalize pre- and post-stimulus to the genomic control, as described above, and then compare the two normalized signals. However, we are really interested in detecting changes between the two conditions, not changes relative to the genomic control. In this case, a more sensitive method to detect the relevant changes between the two conditions would be to directly apply the normalization algorithm described above to the two conditions, i.e. use the pre-stimulus data as the reference condition for the post-stimulus data, and use the post-stimulus data as the reference condition for the pre-stimulus data. This amplifies the differences between the two signals. Then the correspondence between these asymmetric approaches is a measure of the reliability and significance of the detected changes in normalized signal.

3.2.5 Signal Quality Measurement

In the following, we describe a method to quantitatively evaluate the performance of the proposed normalization methods. We define the Signal Quality measure as shown in Figure 3.7. We assume that biologically significant changes between any two arrays will have significantly different signal, but will also have a spatial extent that covers many adjacent probes. To define a set of probes whose signal is significantly changed in two conditions, we use the top 2% of probes sorted by the difference in signal between the two arrays spatially averaged over a window of 147bp. To avoid biasing the results toward one normalization approach, we use the intersection of

the top 2% probes from each approach being compared (say, Group Normalization, Affymetrix MAS5 [69], MAT [66], and quantile normalization [60] of raw data, for which there was a 61% overlap between the top 2% probes using the four methods). These significantly changed probes are indicated by open circles in Figure 3.7. Then the signal power, S, is defined to be the mean square change of the signal on these probes between conditions A and B. The noise power, N, is the mean square change of the signal on entire probes between condition B and a replicate of condition B, as depicted in Figure 3.7.

Using data from [54], three tiling arrays at T=0 (prior to glucose addition) and three tiling arrays at T=60 mins (after glucose addition) were independently normalized against four separate genomic controls using Group Normalization (binary method with four sets of low and high probe ranges as described above, and quantile based method), MAS5, MAT, and quantile normalization of raw data. To make a fair comparison between MAS5 and the other methods, we used a running average window of 20bp to match the 20bp bandwidth in MAS5. Then the Signal Quality, S/N, over the significantly changed probe set in dB was calculated as $10\log_{10}(S/N)$, for all 36 combinations of conditions A, B and replicates. We used these combinations to estimate the mean and standard deviation of the improvement in Signal Quality.

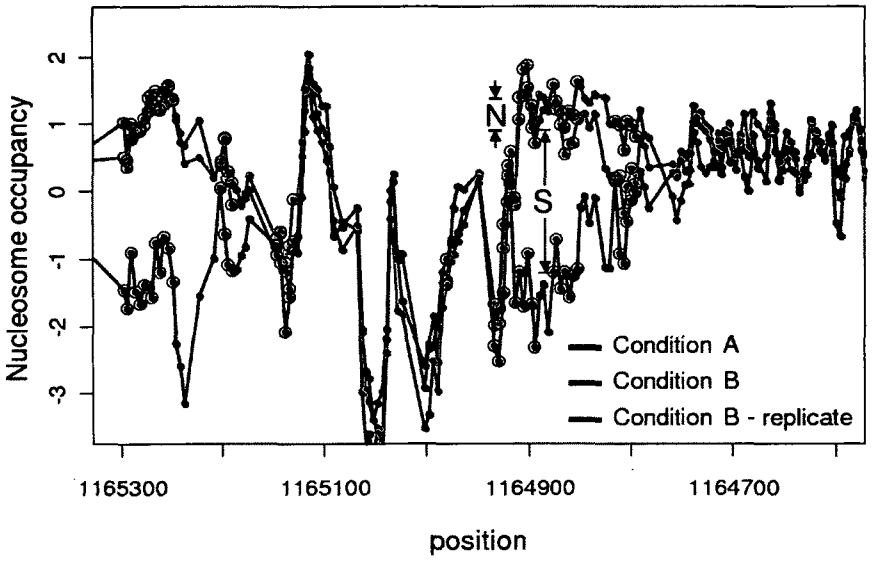


Figure 3.7: Signal Quality measure. Two tiling array signals corresponding to nucleosome occupancy at two different experimental conditions are shown for the *HXT3* locus. We use two conditions and a replicate to determine signal and noise, as follows. In condition A (with glucose), the highlighted region is nucleosome free, and in condition B (no glucose), it is nucleosome bound. S is the difference of the tiling array signal at two different conditions and reflects the signal strength. N is a measure of noise and is estimated by comparing the signal of two replicate microarrays at similar experimental condition. We evaluate S over a set of significantly changed probes (indicated with open circles) and N over all the probes as described in the text. The ratio S/N is a genome wide measure of Signal Quality.

3.2.6 Detecting Enriched Regions in a Spike-in Benchmark Dataset

To compare the performance of the group normalization, with some other existing methods, we applied this method on the Affymetrix and Agilent arrays data in the benchmark spike-in dataset [70]. For Affymetrix arrays, among the methods compared in [70] MAT gives the best performance. To compare our method with MAT, we substituted the probe standardization step in MAT with group normalization and used the same method to detect enriched regions. We used rMAT [71] with the following parameters: dMax = 600, dMerge = 300, nProbesMin = 8, method = "pValue", threshold = 0.0001. For Agilent arrays, among the methods compared in [70] Splitter gives the best performance. To compare our method with Splitter, we substituted the normalization step in Splitter with group normalization and used the same method to detect enriched regions. We used the online implementation of Splitter (http://zlab.bu.edu/splitter) with the following parameters: maxgap=200, minrun=2, mean, Signal cutoff=2.5s.d.

To compare the performance of the methods, we plotted the ROC-like curves similar to [70] and used area under the ROC-like curves as a measure to compare different methods. For a perfect classifier this area is 1, and for random it is near zero.

3.3 Results

We applied the proposed Group Normalization method on published genome-wide nucleosome positioning data in yeast [10, 13, 54]. The joint distribution of probes in the experiment (nucleosome enriched tiling array) and control (genomic hybridization) before and after normalization is depicted in Figure 3.8A. Before normalization, there is a high correlation between signal in the experiment and in the control, which reflects the strong probe effect. This correlation between treatment and control is almost completely removed by our Group Normalization. Figure 3.8B shows normalization results for nucleosome occupancy near the HXT3 promoter before, and 60 minutes after, glucose addition. The array signal changes significantly between the two conditions, with a spatial scale of ~150bp, indicating that nucleosomes at the promoter are removed after glucose addition. Figure 3.8C shows the results for the cross normalization algorithm along the broader HXT locus on chromosome 4. The top row shows the 20bp running average of the raw data and the bottom row shows the results using the cross normalization procedure described above. Cross normalization highlights the regions of differential nucleosome occupancy across this locus much more clearly than can be detected from the raw data. Specifically, it is clear from the cross normalization that while nucleosomes are depleted from HXT3 promoter upon glucose addition, new nucleosomes have been placed at HXT6 and HXT7 promoters. These changes are consistent with the expression level changes for these genes, and with the facts that HXT3 is a low affinity glucose transporter and its

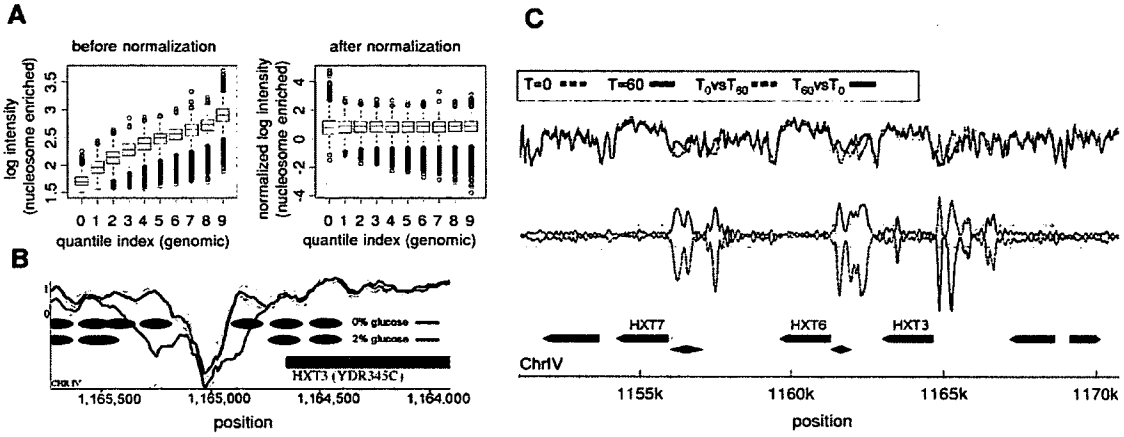


Figure 3.8: Group Normalization results for nucleosome positioning in yeast: (A) probe distribution before (left) and after (right) Group Normalization. (B) Inferred nucleosome pattern at *HXT3* promoter before (blue ovals) and after (red ovals) glucose addition. *HXT3* is upregulated at high glucose levels and repressed at low glucose levels. (C) Differential nucleosome occupancy in yeast in response to glucose addition: cells are grown on glycerol and then 2% glucose is added. Nucleosome positioning is measured before and 60 min after glucose addition [54]. The top curves show the spatially averaged raw tiling array data, at time zero (gray dotted) and t=60 (magenta). The lower plot shows the result of our normalization method. The red curve is the normalized differential nucleosome occupancy for t=60 min compared to t=0 (high values imply increase in nucleosome occupancy in response to glucose). The blue dotted curve is the reverse analysis, comparing t=0 to t=60. The yellow diamonds indicate *ADR1* binding regions from ChIP.

expression is up-regulated upon addition of glucose [72], and that HXT6 and HXT7 are high affinity glucose transporters.

We next applied our Group Normalization method to another nucleosome occupancy dataset, measuring nucleosome occupancy in a histone H3 mutant strain [13]. Figure 3.9 depicts the results of [13], who used Affymetrix Tiling Analysis software (TAS) provided by Affymetrix. We reproduced the nucleosome occupancy profile in the AGE1 locus over the same region shown in Figure 8 of [13], shown in Figure 3.9A.

We also used Group Normalization to process the data for the same region, shown in Figure 3.9B. The most significant change in nucleosome occupancy near AGE1 is that nucleosome binding has been reduced in the histone H3 mutant compared to wild type strain. While this is barely detectable in Figure 3.9A, and somewhat more evident with Group Normalization in Figure 3.9B, cross normalization differentially amplifies these differences in nucleosome occupancy in the two strains, as shown clearly in Figure 3.9C, highlighting the significant difference in occupancy upstream of AGE1.

To quantitatively compare the genome-wide performance of the proposed normalization method with existing methods, we defined a Signal Quality measure (see section 3.2.5), the difference of the signal of two microarrays at two different experimental conditions (Signal) divided by difference of two microarrays at similar condition (Noise). Following [10, 13, 54], who used the Affymetrix MAS5.0 algorithm to process their tiling array data, we compared our Group Normalization method with MAS5.0. We also computed the MAT [66] normalized signal and quantile normalized signal [60] for comparison. We measured the Signal Quality for regions of the genome that are differentially occupied before and after addition of 2% glucose [54] across the whole genome. As shown in Figure 3.10, Signal Quality was significantly improved from 8.8(0.7) dB using MAS5 to 10.5(0.6) dB using binary group normalization, showing 1.7(0.4) dB improvement in Signal Quality compared to MAS5.0 and 1.1(0.8)dB improvement compared to MAT. The numbers in parenthesis are the standard deviation for thirty six different data sets as explained in the section 3.2.5. Also

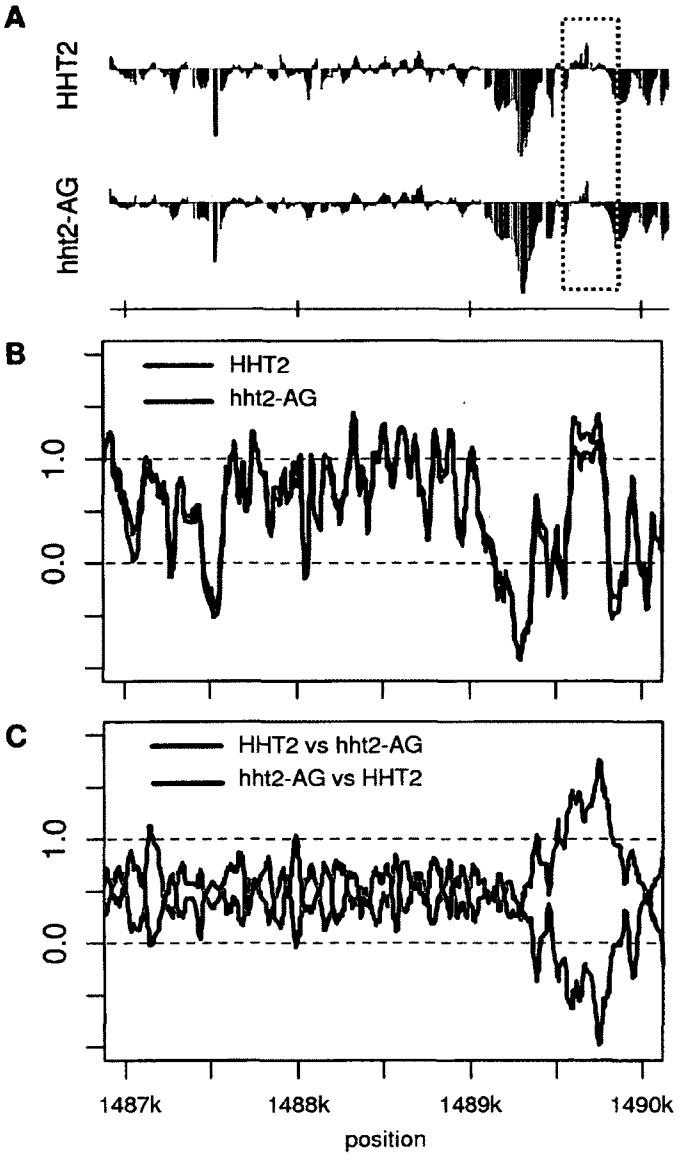


Figure 3.9: Group Normalization results for histone H3 mutant dataset. Nucleosome occupancy in wild type (HHT2) and histone H3 mutant (hht2-AG) near AGE1 on yeast chromosome IV is show for region plotted in Figure 8 of [13]. A) Nucleosome occupancy plots using Affymetrix TAS software as was used by [13]. The dotted box shows the location for the change in nucleosome occupancy. (B) Group Normalization makes it somewhat easier to detect the differentially occupied promoter and clearly identifies the bound regions, but (C) cross normalization more strongly amplifies the differentially occupied region.

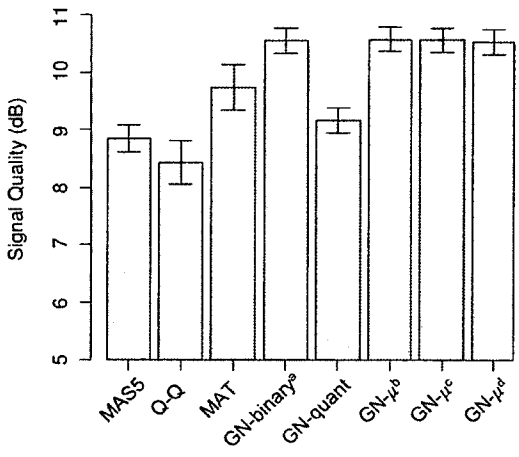


Figure 3.10: Signal Quality comparison of Group Normalization to other methods. We applied different normalization methods to the nucleosome positioning data and measured the Signal Quality using MAS5, quantile normalization (Q-Q), and MAT. Binary Group Normalization (GN-binary) has higher Signal Quality than all other approaches tested. Quantile normalization (GN-quant) outperforms MAS5 and Q-Q but not MAT on this dataset. We also examined the sensitivity of binary Group Normalization to different choices of low and high probe ranges used to estimate μ_{low} and μ_{high}: (μ_{low}, μ_{high})= a: (.10–.40,.60–.90), b: (.05–.50,.80–.95), c: (.10–.50,.50–.90), and d:(.10–.30,.70–.90). All of these choices give virtually identical Signal Quality improvement.

as shown in this figure, using different ranges for low and high probe didn't result in a significant difference in binary group normalization method performance. When we used the alternative quantile-based group normalization method described in section 3.2.1, we found a Signal Quality of 9.2(0.6) dB, so the improvement is 0.4(0.3) dB compared to MAS5 but significantly less than binary method.

The performance of group normalization is not very sensitive to the definition of high and low probe ranges. Considering the probe signal model of Equation (3.1), if the reference set probes are independent of the biological signal, x_i, then for all

reference sets, the observed signal y_i, would have a similar expected distribution to x_i, scaled by A_i and shifted by B_i. Therefore, using different ranges for low and high signal, would give the same normalized signal, except for a constant shift and scale factor, which are functions of the ranges used for low and high signal estimation. To examine the sensitivity of the method to the choice of the low and high probe ranges, we used four different sets of ranges for low and high probes: a) 10%–40%, 60%–90%, b) 5%–50%, 80%–95%, c) 10%–50%, 50%–90%, d) 10%–30%, 70%–90%, and compared the performance of the method using each set of ranges for the nucleosome positioning data, also shown in Figure 3.10. Since as expected the performance was very similar for all the four different ranges, we only used one set of ranges, 10%-40% to define $\mu_{i,low}$ and 60%-90% to define $\mu_{i,high}$ for the spike-in ChIP-chip analysis.

To further evaluate our proposed method against existing methods we applied the group normalization method on the benchmark spike-in dataset [70]. In this dataset, a known amount of DNA from defined cloned regions was spiked in to genomic DNA and performance of different platforms and algorithms was assessed by comparing their ability to accurately recover the spike-in regions. Compared to nucleosome positioning data, where a significant fraction of the probes differ in treatment and control conditions, in this data set the spike in regions only cover about 0.2% of the probes, which is similar to expected ChIP-chip data with limited targets. We compare our method's ability to detect these spike-in regions to alternative methods in Figure 3.11. Group Normalization using either the binary method (GN-binary) or

quantile method (GN-quant) detects more spike regions than Splitter or MAT at the same false positive rate, as shown by the ROC-like curves in Figure 3.11A, defined as in [70]. The area under these curves summarizes the performance of each algorithm, as shown in Figure 3.11B and 3.11C, showing consistent improvement over previous methods.

3.4 Discussion

In this chapter, we have presented a new normalization procedure for genomic datasets. Our approach is based on the idea that genomic data sets have such a large number of sequences or probes that we can estimate the sequence biases of hybridization or sequencing (the dynamic parameters of the probes) from the response of probes across datasets. Because this approach estimates the dynamic parameters of a probe from a group of similar probes, we call our method Group Normalization. We have also described an approach based on this technique which highlights regions of significant signal changes between two experiments (Cross Normalization). We have shown that these normalization procedures can significantly improve the signal quality relative to the existing normalization methods. We have shown that Group Normalization improves signal to noise in nucleosome positioning datasets and can more accurately identify spiked-in regions in the benchmark data of [70].

While signal quality is a global measure which shows the benefit of Group Normal-

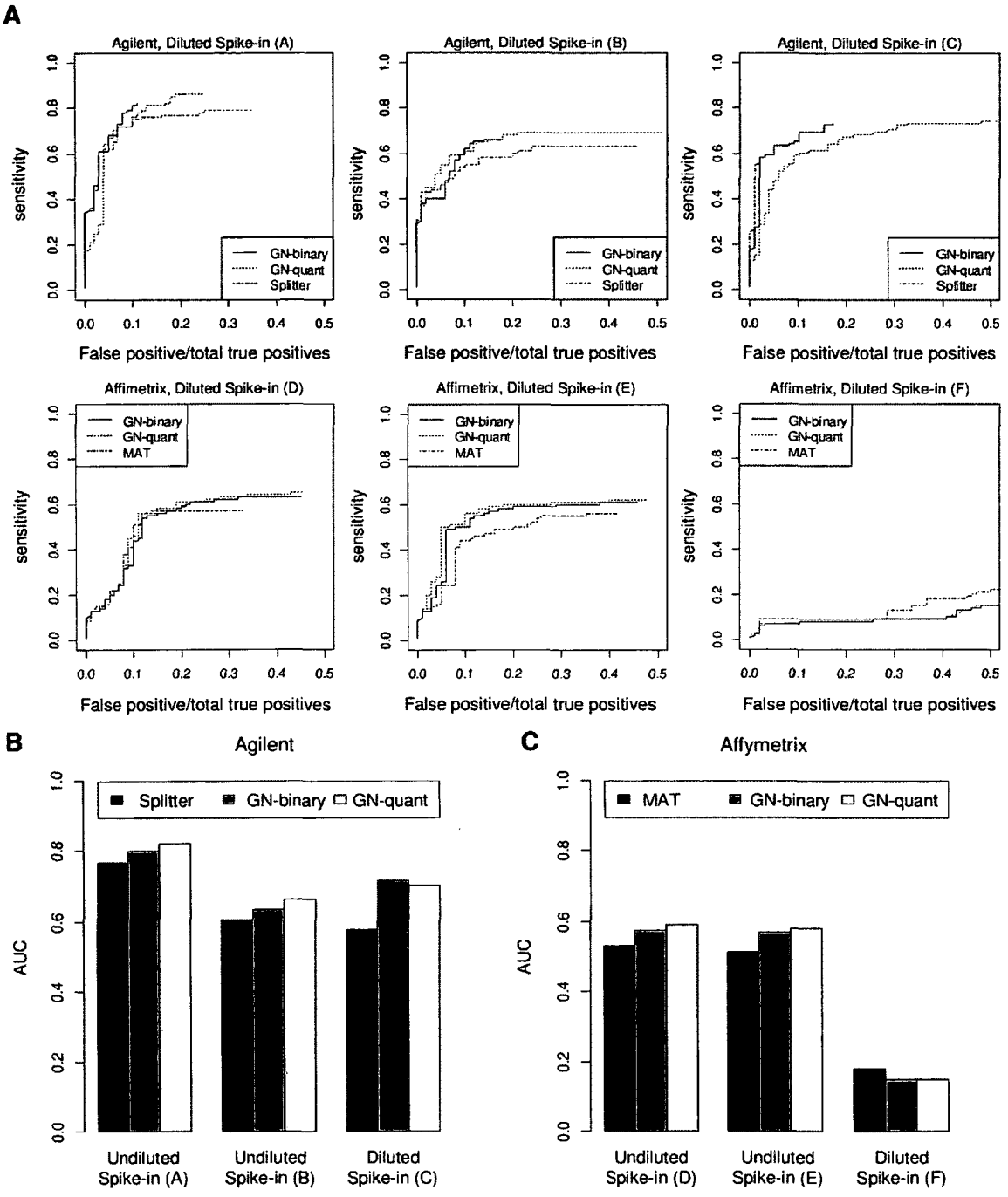


Figure 3.11: Comparison with spike-in benchmark data of [70]. A) We compare ROC-like curves for different platforms and algorithms: Splitter, which had the best performance on Agilent data, and MAT, which had the best performance on Affymetrix data. Area under the ROC-like curve (AUC) is shown for B) Agilent and C) Affymetrix datasets. Except for the diluted Affymetrix spike-in data, which had poor performance with all methods, Group Normalization (both GN-binary and GN-quant) consistently performs better than previous methods, and has a higher sensitivity to recover spike-in regions at the same false positive rate.

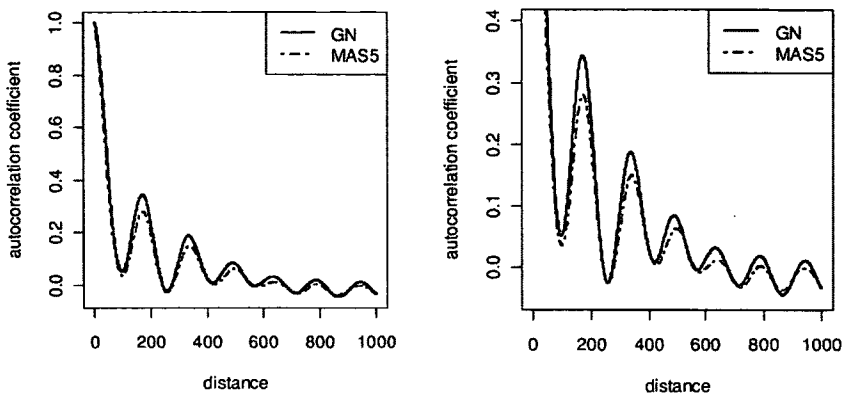


Figure 3.12: Comparison of autocorrelation of normalized nucleosome occupancy using Group Normalization and MAS5 algorithms for the [10] data. Group normalization shows a slightly higher recurrence in nucleosome occupancy signal due to the periodic packing of nucleosomes.

ization compared to other approaches, in some cases, the biology under consideration can also show how Group Normalization improves the analysis of the data. For nucleosome positioning, the known spatial scale of DNA bound to the nucleosome constrains the signal to vary on a scale of 150bp. We can use genome wide nucleosome occupancy data to construct the autocorrelation of the nucleosome bound DNA signal for data that has been normalized using Group Normalization or other approaches. Because of nucleosome packing, we expect this autocorrelation function to exhibit periodicity on a scale set by the 150bp nucleosome bound DNA plus linker DNA. As shown in Figure 3.12, the autocorrelation from group normalized data shows a stronger recurrence at 170bp in the autocorrelation function compared to MAS5.

A possible extension of this model would incorporate an estimate of the variance of the probe signal from the reference set, and use that as a measure of the reliability

of a probe. MAT uses a similar approach when combining the signal of neighboring probes.

While in this chapter we have only presented results on tiling microarray data, in principle, other high-throughput technologies could benefit from a model-independent normalization approach similar to Group Normalization. Our development of the Group Normalization procedure was motivated by the observation that tiling array probes exhibit widely varying hybridization efficiencies, presumably due to non-uniform variations in the local sequence properties of the probes. We normalized these varying hybridization efficiencies by finding a large set of similarly responding probes in a reference condition. Most other genomic technologies suffer from analogous sequence specific effects on the assay efficiency. These could range from sequence dependent shearing rates, endonuclease sequence cleavage preferences, or sequence specific priming efficiencies in the case of massively parallel sequencing assays (RNA-seq or Chip-seq). Because we do not try to explicitly model the sequence specificity of the assay, but instead infer (estimate) sequence specific probe effects from a reference set of probes, our proposed approach should be useful in these cases as well.

Chapter 4

Context-Based Model for Nucleosome Positioning

4.1 Introduction

Nucleosomes occupy specific positions in most of the genes in yeast [10, 11, 73]. Typically, a nucleosome-free region (NFR) is formed at gene promoter region, which is usually enriched in transcription factor binding sites. The NFR is followed by a well-positioned nucleosome that covers the transcription start site (TSS) and usually incorporates a histone H2A.Z variant [45]. Formation of the NFR is believed to facilitate binding of transcription factors (TFs) to their binding sites. Some DNA sequences have lower affinity to bind histones and/or are less flexible to form nucleosomes [8]. NFR formation near TSS and transcription termination sites (TTS) appears *in vitro*

as well. Hence, the relatively weak nucleosome affinity in the NFR near the TSS is thought to be encoded in the DNA sequence of these regions [11]. However, the precise positioning of the nucleosomes *in vivo* is affected by ATP-dependent nucleosome remodeling mechanisms [14]. This explains the difference between nucleosome occupancy profile *in vivo* and *in vitro* (Figure 2.4).

Recent work [11, 38] has modeled the sequence dependence of nucleosome positioning using a combination of dinucleotide frequencies with periodicity that is usually modeled rigidly (10.5 bp) [11] or with varying periodicity (2, 4, 8, 16, 32, 64, 128bp) [38]. But we expect that as DNA wraps around nucleosome, some flexibility in this rigid periodicity could be easily tolerated. Moreover, dinucleotide frequencies do not represent all the relevant structural properties or physical interactions between DNA and nucleosomes or other structures. In this chapter and in the next two chapters, we describe three novel sequence models we designed to more accurately predict nucleosome positioning. First we describe the *simple context-based* model which is inspired by context-based probability modeling used in arithmetic coding [74, 75]. Then we describe a *phase-dependent context-based* model which improves context modeling for the *context-based model* based on a roll-twist-and-tilt model for DNA bending. Then, in chapter 5 we describe a novel method for robust estimation of k-mer frequencies, and based on that we develop the *gapped k-mer support vector machine (GSVM)* model, which is described in chapter 6. Finally we compare the performance of these methods by directly assessing their ability to predict the *in vivo*

and *in vitro* nucleosome positioning data.

4.2 Simple Context-Based Sequence Model

When wrapped around the core histone octamer, the structural conformation of DNA deviates from its native super helical form because of site specific DNA-protein interactions [76]. A direct model for sequence features that affect these interactions would necessarily be extremely sophisticated; however, we propose a data driven model that is flexible enough to capture the most significant of these interactions and learn sequence features from nucleosome positioning data.

We build two probability models for DNA sequences, one for the nucleosome bound regions and one for nucleosome free regions. The probability of each base pair occurring in nucleosome bound vs. unbound regions is estimated from its local sequence context (neighboring base pairs in both directions). This is done by counting the number of times each of the four possible bases appears in the training data for each context model in each of the two classes (bound/free). This is shown in Figure 4.1. The number of neighboring nucleotides forming the context determines how accurately we can estimate the probabilities. Larger contexts can describe more specific combinations of bases and their periodicity; however, the problem with too many context bases is that we do not have enough training data to robustly estimate the distribution within each context. To overcome this, we dynamically set the con-

text size (maximum of 12 bp) for each context model to reach a minimum number of samples (e.g. 100) in the training set; this guarantees robust probability estimation and prevents over-fitting.

To predict whether a given sequence of length 147 bp would belong to nucleosome bound class, the average number of bits required to code the sequence using nucleosome bound model (equivalently, the negative log likelihood) is calculated as follows:

$$H_{bound} = -\frac{1}{N}\sum_{i=1}^{N}\log_2(P_{bound}(b_i|\text{context}_i)) \tag{4.1}$$

where (N=147) is the sequence length. Similarly, the average number of bits is calculated for the un-bound model as follows:

$$H_{unbound} = -\frac{1}{N}\sum_{i=1}^{N}\log_2(P_{unbound}(b_i|\text{context}_i)). \tag{4.2}$$

The difference of the above two gives the model score:

$$\Delta_H = H_{unbound} - H_{bound} \tag{4.3}$$

The higher the model score, the more likely the sequence is nucleosome bound. We have tested the performance of this model to classify regions that are unambiguously nucleosome bound or nucleosome free and quantify the models performance by the area under its ROC curve (AUC). Using data from [10](Figure 4.2B) we find high predictive accuracy: AUC=0.94. In another dataset we used the model to discriminate highly nucleosome enriched probes from nucleosome depleted probes in [77] (Figure 4.2C), with AUC=0.9. Finally, a similar dataset was used in [38] and our

Nucleosome-bound

Sequence	Frequency f	Probability p	# of bits -log(p)
AAACA**A**GTGCA AAACA**A**GTGCA ... AAACA**A**GTGCA	6	P_{bound} (**A** | AAACA-GTGCA)= 6/21	1.81
AAACA**C**GTGCA AAACA**C**GTGCA ... AAACA**C**GTGCA	7	P_{bound} (**C** | AAACA-GTGCA)= 7/21	1.58
AAACA**G**GTGCA AAACA**G**GTGCA ... AAACA**G**GTGCA	5	P_{bound} (**G** | AAACA-GTGCA)= 5/21	2.07
AAACA**T**GTGCA AAACA**T**GTGCA AAACA**T**GTGCA	3	P_{bound} (**T** | AAACA-GTGCA)= 3/21	2.81

Nucleosome-unbound

AAACA**A**GTGCA AAACA**A**GTGCA AAACA**A**GTGCA	3	$P_{unbound}$ (**A** | AAACA-GTGCA)= 3/22	2.87
AAACA**C**GTGCA AAACA**C**GTGCA ... AAACA**C**GTGCA	5	$P_{unbound}$ (**C** | AAACA-GTGCA)= 5/22	2.14
AAACA**G**GTGCA	1	$P_{unbound}$ (**G** | AAACA-GTGCA)= 1/22	4.46
AAACA**T**GTGCA AAACA**T**GTGCA ... AAACA**T**GTGCA	13	$P_{unbound}$ (**T** | AAACA-GTGCA)= 13/21	0.76

Figure 4.1: Context-dependent probability modeling: The probability of each base in each class (bound/unbound) is estimated based on the number of samples in the training set that have the same context sequence. In the above example, the probability of the central base to be T given that the neighboring bases are AAACA-GTGCA is 3/21=0.14 in the bound class versus 13/21=0.62 in the unbound class.

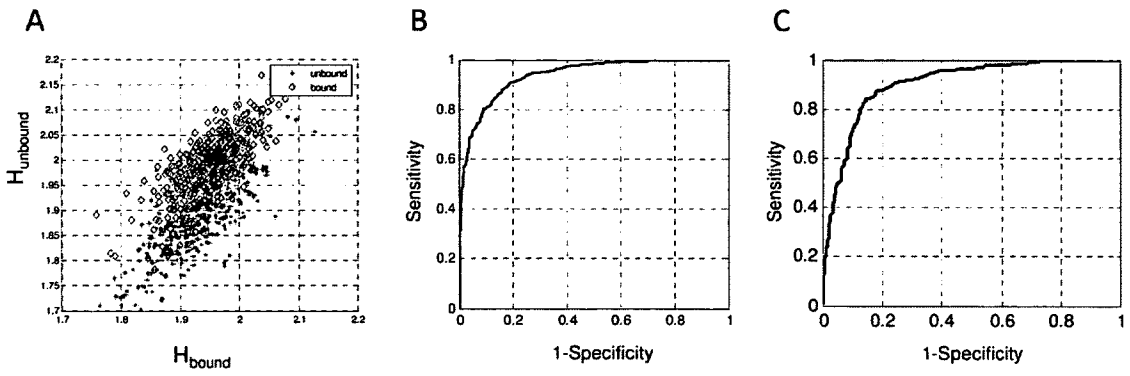


Figure 4.2: Context-based model classification results: A) Scatter plot for $H_{unbound}$ vs H_{bound} calculated as given in equations (4.1),(4.2) for nucleosome bound sequences (blue diamonds) and nucleosome free sequences (red stars). B) ROC curve for classification performance of unambiguously nucleosome bound and nucleosome free regions from [10]; AUC=0.94. C) ROC curve for discriminating highly nucleosome enriched probes from nucleosome depleted probes in [77], AUC=0.9.

proposed context-based model outperforms N-score model [38], AUC=0.88, Segal's original model [8], AUC=0.61, and Peckham's model [39], AUC=0.78.

4.3 Phase-Dependent Context-Based Model

Different DNA sequences have different flexibilities and preferred curvature. In particular it has been reported that certain dinucleotides appear periodically in nucleosome bound sequences [8]. The period of these periodic features is estimated to be around 10.5 bp and coincides with the DNA double helix period. Thus, it appears that some sequences mostly appear on the outer surface of the nucleosomes while other sequences preferentially are in direct contact with the core histones. To form a nucleosome, the DNA molecule needs to make sharp bends [78]. This total curvature

is distributed over individual dinucleotides. It has been shown that different dinucleotides have different stiffness and different preferred conformations [79, 80], and elastic models, such as the model proposed in [81] can be used to model DNA bending energies. However, in a nucleosome, forces and torques exerted by the histones core cause the DNA conformation to deviate significantly from the predicted least elastic energy conformation as shown in [76, 82]. Therefore models merely based on DNA elasticity that ignore the DNA-protein interactions may not accurately predict preferred nucleosome positions. Our approach is to employ a simple shear free elastic model to predict the phase for each dinucleotide in a bent DNA molecule from the stiffness and preferred roll, tilt and twist angles of dinucleotides. This determines the position of dinucleotides relative to the core histones (i.e. whether a dinucleotide is facing towards the core histone or it is on the outside surface of the nucleosome). Then we integrate the phase information into the context-based model to better estimate the probability distributions for nucleosome bound and nucleosome free models.

4.3.1 Phase Estimation

Similar to [76,81,82,86] we model each base-step as a rigid block (brick). Figure 4.3 shows the brick model representation of the NCP147 nucleosome [83]. Using the brick model allows us to fully identify the position and orientation of each base-step in 3-dimensional space by having the position and orientation of the previous base-step plus six additional quantities that determine the position and orientation of a brick

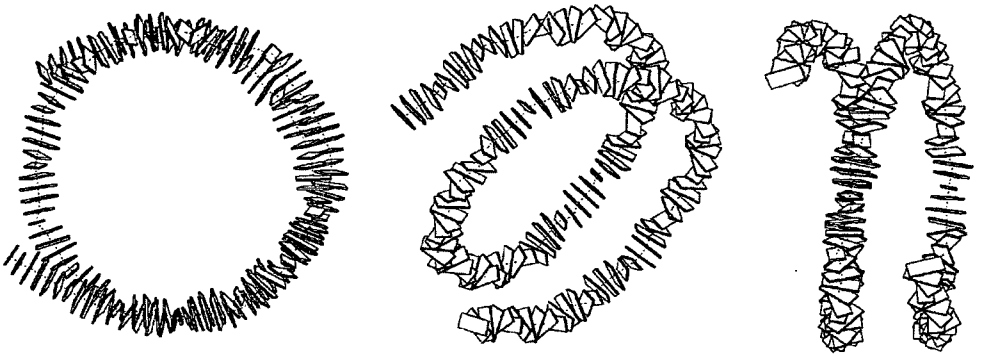


Figure 4.3: Brick representation of NCP147 nuclesome [83], from three different angles. Parameters estimated from crystal structure and figures plotted using 3DNA software [84, 85]

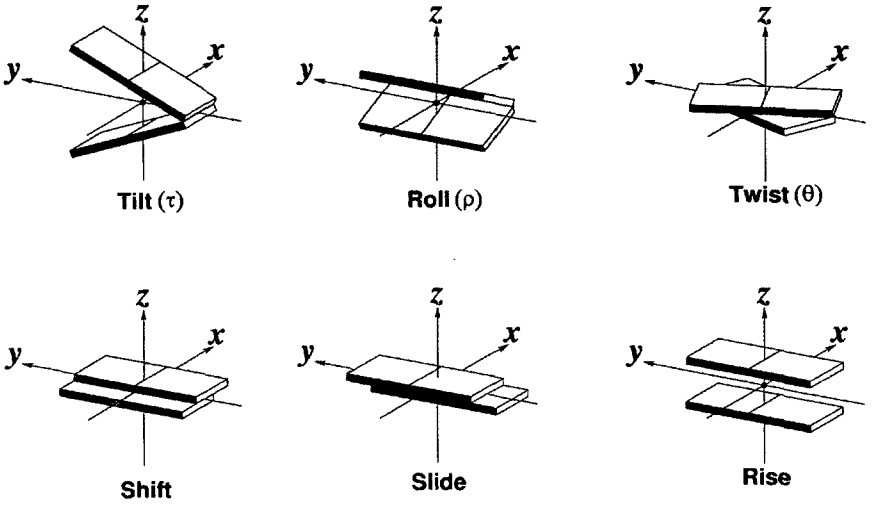


Figure 4.4: The three translational and three rotational axes for pair of adjacent base pairs [78]. Diagrams generated by 3DNA software [84, 85]

relative to the previous brick: three translocations (shift, slide, and rise) and three rotations (roll, tilt and twist angles). These are depicted in Figure 4.4. We model the total bending of the DNA molecule as the sum of the contributions of roll (ρ) and tilt (τ) angles of each base-step multiplied by the cosine and sine of the phase (φ_i) of that base-step:

$$bend_i \cong \rho_i Cos(\varphi_i) + \tau_i Sin(\varphi_i) \tag{4.4}$$

Phase (φ_i) is the accumulated twist angle (θ_i) and determines how the local roll or tilt angle will contribute to the DNA bending around the nucleosome. We estimate the elastic free energy of bending as follows:

$$E(s_j) = \sum_{i=1}^{146} E_i = \sum_{i=1}^{146} \left\{ C_{1i}^2(\rho_i - \rho_{0i})^2 + C_{2i}^2(\tau_i - \tau_{0i})^2 + C_{3i}^2(\varphi_i - \varphi_{i-1} - \theta_{0i})^2 \right\} \tag{4.5}$$

where E_i is the free energy contribution from base-step at position i, and consists of three terms to model the free energy changes for deviations of roll, tilt, and twist angles from their preferred values ρ_{0i}, τ_{0i}, and θ_{0i} (obtained from [80]) weighted by corresponding stiffness parameters C_{1i}, C_{2i} and C_{3i} (obtained from [79]). To find the phasing of each base of a 147 bp long fragment of DNA in nucleosome bound conformation, we minimize the free energy given in equation (4.5) with the constraint that the fragment has to wrap around nucleosome twice; i.e. $\sum bend_i = 4\pi$. To solve this optimization problem efficiently, we used a two step dynamic programming approach as shown in Figure 4.5: First we initialize $bend_i$ by uniformly distributing the total bend to every base-step. Then we keep the bend angles fixed and optimize twist angles (θ_i's) using dynamic programming. For this, we define the recursive

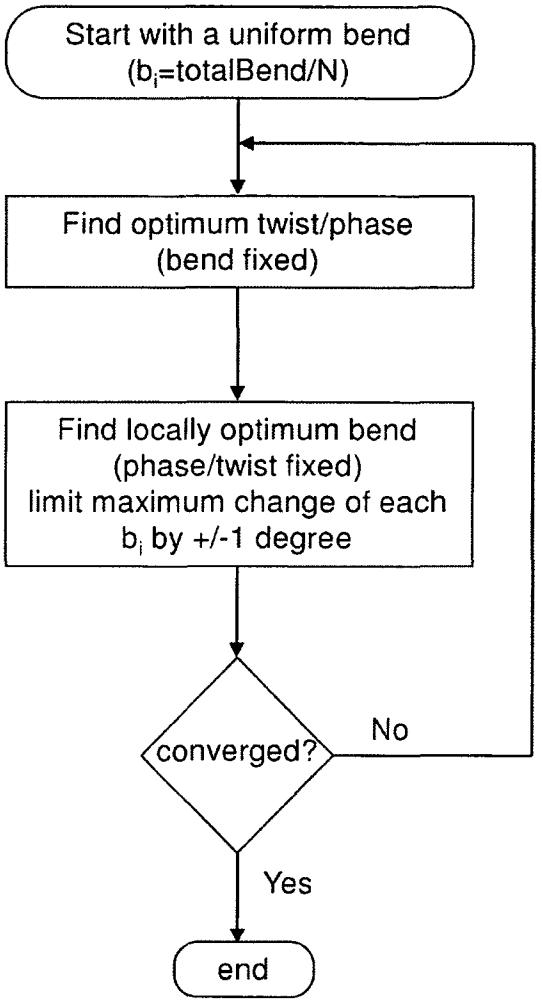


Figure 4.5: Overview of the two step iterative optimization algorithm for minimum free elastic energy estimation of twist angles and bend distribution.

penalty function as follows:

$$P(n,\varphi) = \min_{\theta_{\min} \le \theta_n \le \theta_{\max}} P(n-1, \varphi - \theta_n) + f(\theta_n, b_n; \varphi - \theta_n) \tag{4.6}$$

where $f(\theta_n, b_n; \varphi)$ is the local energetic cost of the n'th base step to take bending of b_n degrees, with twist of θ and phase of φ as defined in equaltion (4.8) below. $P(n,\varphi)$ is the minimum partial energetic cost for base steps $0,\ldots,n$ having bend angles $b_0,\ldots,b_n$ with ending phase equal to φ. The initial condition for this recursive equation is given by $P(n,\varphi) = 0$ for $n < 0$ and φ is cyclic, meaning that $2\pi + \varphi \equiv \varphi$. This recursive form can be efficiently solved using the dynamic programming algorithm [87]. Then we fix the twist angles and find the optimum bend angles. For this we define a similar recursive cost function as follows:

$$P(n,B) = \min_{b_n - \Delta \le b_n \le b_n + \Delta} P(n-1, B - b_n) + f(\theta_n, b_n; \varphi - \theta_n) \tag{4.7}$$

where $P(n,B)$ is the minimum energetic cost for base steps $0,\ldots,n$ having phases $\varphi_0,\ldots,\varphi_n$ forming total bend equal to B. The initial condition for this recursive equation is given by $P(n < 0, B = 0) = 0$, $P(n < 0, B \neq 0) = \infty$ and $P(n, B > maxTotalBend) = P(n, B < minTotalBend) = \infty$. This recursive form can be efficiently solved using dynamic programming technique. We used the following discretization of parameters: $20° \le \theta_i \le 50°$, resolution(step size)=1°, $-5° \le bend_i \le 15°$, resolution=0.5°, $0° \le \varphi < 360°$, resolution=1°, $-10° \le totalBend \le 730°$, resolution=0.5°. Finally we define local bending energetic cost for a base step, $f(\theta, b; \varphi)$, based on the base step parameters (mean and variance of roll, tilt and

twist angles) as follows:

$$f(\theta, b; \varphi) = \log(\sigma_b) + \frac{(b - \mu_b)^2}{2\sigma_b^2} + \log(\sigma_\theta) + \frac{(\theta - \mu_\theta)^2}{2\sigma_\theta^2} \tag{4.8}$$

where $\mu_\theta, \sigma_\theta^2$ are the mean and variance for twist angle, and $\mu_b = \mu_\rho cos(\varphi) + \mu_\tau sin(\varphi)$, $\sigma_b^2 = \sigma_\rho^2 cos^2(\varphi) + \sigma_\tau^2 sin^2(\varphi)$ are the mean and variance for bend angle.

4.3.2 Phase Estimation Results for NCP147

To evaluate the performance of the DNA elastic model and the optimization algorithm, we applied our model to the NCP147 nucleosome sequence [83] to predict the phase. Results are depicted in Figure 4.6. It can be observed that the estimated phase using the elastic model is very close to the experimental crystal structure data.

We also evaluated the sensitivity of the estimated phase to the total bend angle by estimating the phase for different values of the total bend $(2\pi, 3\pi, 4\pi, 5\pi, 6\pi)$. Results are shown in Figure 4.7. It can be observed that the estimated phase is not very sensitive to the total bend angle. By increasing the total bend (over-stretching), each base step is stretched more to yield higher bend, however the phasing remains mostly unchanged as the nucleotides that were facing towards the core remain inside and those on the outside surface remain outside.

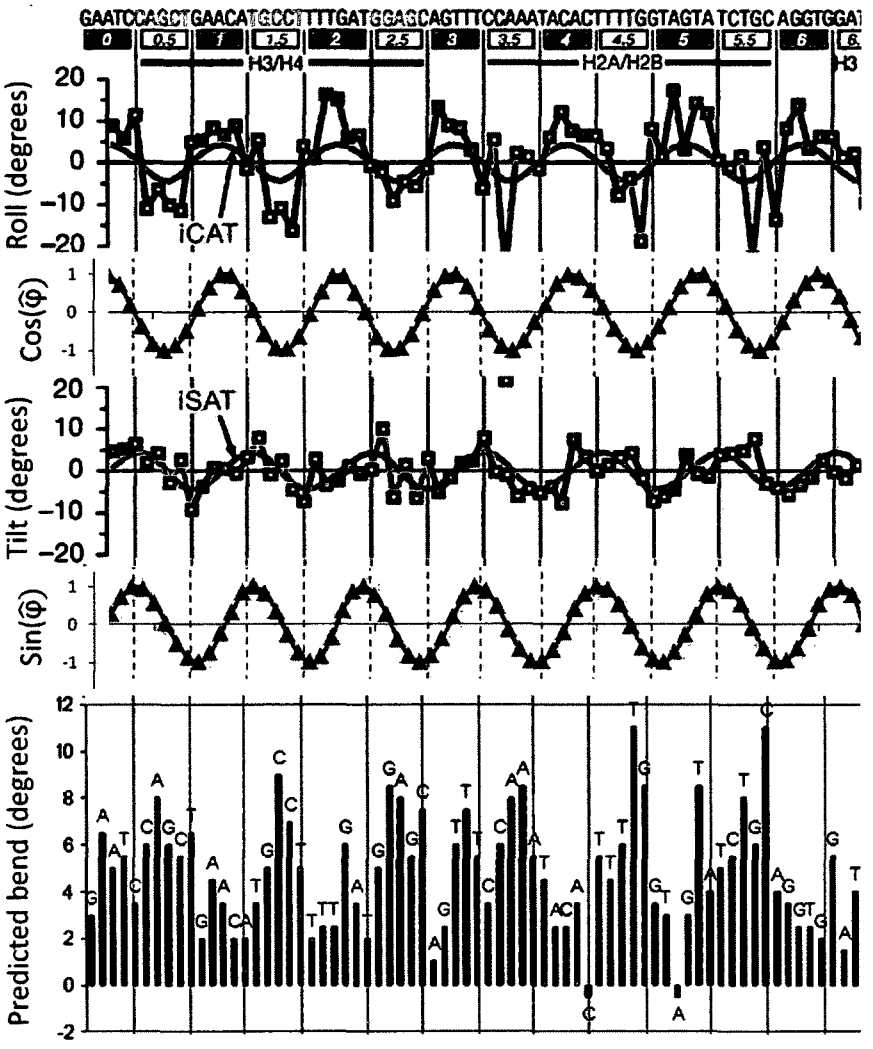


Figure 4.6: Comparison of elastic model prediction and experimental data for NCP147: The top curve is the roll angles inferred from crystal structure of NCP147 nucleosome [83]; Cosine (iCAT) of the phase is shown in red on the roll angle plot on top. Cosine of the estimated phase (model prediction) is plotted in blue on the second plot. Similarly, inferred tilt angles and Sine of the phase (iSAT) is plotted on the third plot, and Sine of the estimated phase (magenta) on the fourth plot. The estimated bending (blue) of the NCP147 sequence by our model is plotted on the bottom

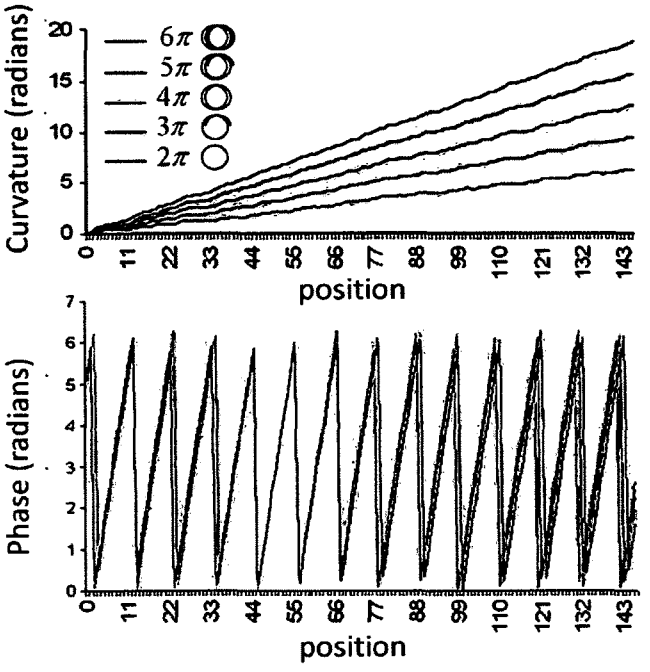


Figure 4.7: Sensitivity of estimated phase to total bend: We estimated the phase and bend angles for NCP147 sequence [83] with various amounts of total bend $(2\pi, 3\pi, 4\pi, 5\pi, 6\pi)$. The top curve shows the accumulative bend angles along the sequence for different total bend angles. The bottom plot shows the estimated phase angles along the sequence for different total bend angles.

4.3.3 Integrating Phase Model with Context-Based Sequence Model

As we described in the previous section, we can estimate the phase for each nucleotide in a nucleosome bound DNA sequence using the elastic model. The estimated phase gives the information about the position of each base in the nucleosome with respect to the core histones. To integrate this information with the context-based model described earlier, we bin the estimated phase values and for each bin, we estimate the probability distributions for bases independently. We start the bin number from the center, so that the base near the center of the nucleosome will be placed in bin zero, and all other bins are numbered relative to that. This is shown in Figure 4.8 for two different binning intervals π and $\pi/2$. A comparison of performance of *simple context-based model* and *phase-dependent context-based model* will be given in section 6.3.2.

While we have shown that the incorporation of phase information through intrinsic flexibility and curvature improves the ability of our models to predict nucleosome positions, these more complicated models require significantly more training data to accurately specify the DNA sequence-histone interactions. To circumvent some of the limitations on context size imposed by the finite amounts of training data, we next developed an alternative approach to the context-based-model for describing the DNA sequence-histone interactions based on k-mer sequence features. As described

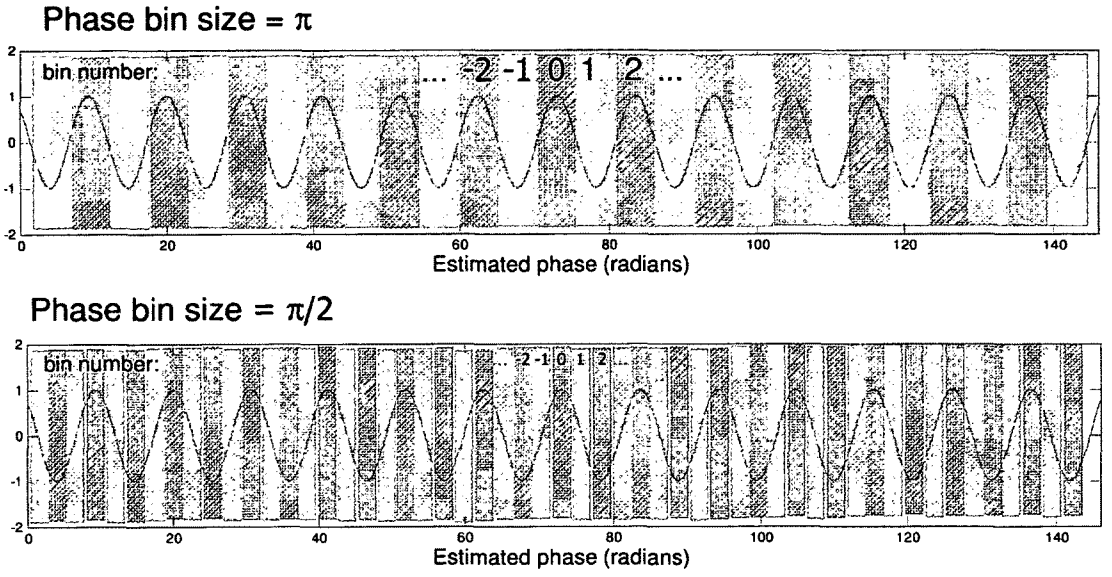


Figure 4.8: Phase binning: The estimated phase is binned (top figure: bin size of π and bottom: $\pi/2$) and bins are numbered starting zero at the center (for the nucleosome dyad axis). Each phase bin will be handled as an independent context model in the conext based sequence model.

in the next section, this k-mer based model is designed to be more flexible in its description on sequence features and is more robust to overfitting the training data.

Chapter 5

Robust k-mer Frequency Estimation Using Gapped k-mers

Oligomers of fixed length, k, commonly known as k-mers, are often used as fundamental elements in the description of DNA sequence features of diverse biological function. In [39] an SVM classifier based on k-mers of length $k \leq 6$ is used to discriminate nucleosome bound and nucleosome free sequences. k-mers are very useful as general sequence features because they constitute a complete and unbiased feature set, and do not require parameterization based on incomplete knowledge of biological mechanisms. However, a fundamental limitation in the use of k-mers as sequence features is that as k is increased, larger spatial correlations in DNA sequence elements can be described, but the frequency of observing any specific k-mer becomes very small, and rapidly approaches a sparse matrix of binary counts. Thus any statistical

learning approach using k-mers will be susceptible to noisy estimation of k-mer frequencies once k becomes large. Because all molecular DNA interactions have limited spatial extent, gapped k-mers often carry the relevant biological signal.

In this chapter, we describe a novel method to use gapped k-mer counts to more robustly estimate ungapped k-mer frequencies, by deriving an equation for the minimum mean square error (MMSE) estimate of k-mer frequencies given gapped k-mer frequencies. Then in chapter 6, we describe a sequence similarity score (gscore) based on the robust k-mer frequency estimation method. We use the proposed sequence similarity score as a kernel for a support vector machine (SVM) to build a general sequence classifier. We will use this classifier to discriminate nucleosome bound and nucleosome free sequences.

5.1 Introduction

Most analysis of polymeric biomolecules (e.g. protein, RNA, or DNA) at some point requires a model mapping polymer sequence features to functional molecular structures. In the context of DNA sequence properties, sequence similarity via sequence alignment, alignment to known repeat elements, and CpG islands are well known examples. In protein functional and structural studies, amino acid patterns are frequently mapped to structural or functional motifs (such as leucine zippers or phosporylation sites). These descriptions have been fairly successful. However, when

modeling DNA-protein interactions, a core process in transcriptional regulation, there is less consensus on what to use as the best description of a DNA sequence (a binding site) bound by a protein (a transcription factor). Many models of this process use a position weight matrix or PWM to describe the DNA binding site [1, 88, 89]. Other approaches use oligomers of fixed length, k, commonly known as k-mers, to describe the DNA binding sites. Using k-mers has the distinct advantage that they reflect the discrete space of all possible DNA molecules of length k, while the space of all possible PWMs is continuous and large. We were initially motivated to generalize k-mer methods when our machine learning algorithms based on k-mer frequencies were found to be effective at predicting enhancers and modeling transcription factor binding sites [58]. Other examples of successful bioinformatic applications based on k-mers as sequence features include sequence homology [90, 91], protein homology [92], and predictions of cis-regulatory modules [93], transcription factor binding sites [94, 95], transcription initiation sites [96, 97], and splice sites [98, 99].

When using k-mers, larger k's will resolve larger binding sites and more accurately reflect biological function. For example, some transcription factors (such as ABF1) have relatively long binding sites that cannot be completely represented by short k-mers. So longer k-mers capture more relevant information; however, there is a limitation on the maximum length k which can be effectively used in statistical algorithms. Because longer k-mers are more sparsely populated in any finite training sequence set, there is a maximum length k for which the k-mer frequencies can

be robustly estimated. Thus in practice, a k is chosen which is a tradeoff between resolving features and robust estimation of their frequencies. To overcome the finite training set size problem, one approach is to employ gapped k-mer frequencies. A gapped k-mer has a length ℓ, and a number of informative columns within that ℓ-mer, k, which reflects the base pairs which actually affect the strength of the TF-DNA binding interaction. We have found that using gapped k-mers can substantially improve the reliability of the k-mer frequency estimation for a finite genomic training set, because while k-mers become sparsely populated, gapped k-mers will still have many instances in the training set, and thus their frequencies can be more reliably estimated. We use the observed gapped k-mer frequency distribution for all gapped k-mers to estimate the ungapped ℓ-mer frequencies, which are sparsely populated. Mathematically we formulate this as the minimum mean square error estimate for the ℓ-mer frequencies given the frequencies for all gapped k-mers. We derive the matrix, W, mapping between these two spaces. A closed form for this matrix is obtained by studying combinatorial properties of the incidence matrix.

This chapter is organized as following: in section (5.2) we formulate the ℓ-mer frequency estimation as a linear optimization problem and show that it can be solved by finding an eigendecomposition for $AA^{\top}$, where A is the incidence matrix. In section (5.3), we provide the necessary preliminary information and some definitions needed for the eigendecomposition and the subsequent proof. In section (5.3.4) we define function ν and prove some identities that enables us to find an eigendecomposition

for matrix AA^{T} in section (5.4). Then in section (5.5) we show that matrix W is the Moore-Penrose pseudoinverse of matrix A. As a consequence, in section (5.6) we find a basis for the row space of matrix A that can further explain the rank and some other features of this matrix. Finally, in section (5.7), we give a summary of the work in this chapter.

5.2 Problem Statement

Given the frequencies for gapped k-mers, we want to estimate the frequencies for any given sequence of length ℓ.

Definition $U = \{u_j\}, 1 \leq j \leq N = b^\ell$ is the set of all different sequences of length ℓ over the alphabet $\{0, 1, \cdots, b-1\}$.

Definition $V = \{v_i\}, 1 \leq i \leq M = \binom{\ell}{k} b^k$ is the set of all gapped k-mers of length ℓ with k known bits and $\ell - k$ gaps.

Definition Matrix $A_{M*N} = [a_{i,j}]$ is a binary matrix defined as the following:

$$a_{i,j} = \begin{cases} 1 & \text{if } v_i \text{ matches } u_j \\ 0 & \text{otherwise} \end{cases}$$

Definition $\mathbf{x}$ is a vector of length N, where x_j is the count for u_j.

Definition $\mathbf{y}$ is vector of length M, where y_i is the count for v_i.

Given the above definitions, we can write the mapping from ℓ-mer counts to gapped k-mer counts as:

$$\mathbf{y} = A\mathbf{x} \tag{5.1}$$

Assuming that x_j's are independent and have the following joint normal distribution:

$$F(\mathbf{x}) = (2\pi)^{-\frac{N}{2}} |\Sigma_{\mathbf{x}}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^\top \Sigma_{\mathbf{x}}^{-1}(\mathbf{x}-\overline{\mathbf{x}})} \tag{5.2}$$

where $\Sigma_{\mathbf{x}} = \sigma^2 I_N$ is a diagonal matrix with constant elements on the diagonal and $\overline{\mathbf{x}}$ a constant vector.[1] We want to find $\mathbf{x}$ that maximizes (5.2) subject to the constraints given by (5.1). The MMSE estimate for $\mathbf{x}$ is given below. Before this, we mention that since $AA^\top$ is a positive semidefinite matrix, it admits the eigendecomposition $AA^\top = Q\Lambda Q^\top$ where the matrix Λ is a diagonal matrix having nonzero eigenvalues of $AA^\top$ on its diagonal and the columns of Q are normalized orthogonal eigenvectors ordered similarly. It is obvious that $Q^\top Q = I$ and it is not hard to prove that $QQ^\top A = A$. The proof is similar to that of Proposition 5.4.2(iii) given later in this chapter.

Theorem 5.2.1 *Suppose that the matrices A, Q and Λ are defined as above. Then the MMSE estimate for $\mathbf{x}$ is given by $\mathbf{x}_{MMSE} = W\mathbf{y}$, where W_{N*M} can be written as the following:*

$$W = A^\top Q \Lambda^{-1} Q^\top \tag{5.3}$$

[1] A more advanced model might have Σ still diagonal but with elements scaling as $\bar{x}$.

Proof We use the technique of Lagrange multipliers to solve the above constraint optimization problem. We will find the $\mathbf{x}$ that maximizes the logarithm of $F(\mathbf{x})$ and also satisfies 5.1. Applying the Lagrange multiplier theorem we have:

$$\mathbf{x} - \overline{\mathbf{x}} = A^{\top}\lambda. \tag{5.4}$$

Reordering (5.4) and applying (5.1) we obtain:

$$\mathbf{y} = AA^{\top}\lambda + A\overline{\mathbf{x}}. \tag{5.5}$$

Now, consider the following eigendecomposition for $AA^{\top}$:

$$AA^{\top} = Q\Lambda Q^{\top}. \tag{5.6}$$

Multiplying both sides of (5.5) by $A^{\top}Q\Lambda^{-1}Q^{\top}$ and applying (5.6) we obtain:

$$\begin{aligned} A^{\top}Q\Lambda^{-1}Q^{\top}(\mathbf{y} - A\overline{\mathbf{x}}) &= A^{\top}Q\Lambda^{-1}Q^{\top}AA^{\top}\lambda \\ &= A^{\top}\lambda. \end{aligned} \tag{5.7}$$

Reordering (5.7) and applying (5.4) we obtain:

$$\mathbf{x}_{MMSE} = W(\mathbf{y} - A\overline{\mathbf{x}}) + \overline{\mathbf{x}}. \tag{5.8}$$

In our case $\overline{\mathbf{x}}$ is a constant vector. Then, since the sum of the elements in rows of matrix WA is equal to 1, equation (5.8) can be simplified to

$$\mathbf{x}_{MMSE} = W\mathbf{y} = A^{\top}Q\Lambda_0^{-1}Q^{\top}\mathbf{y}. \tag{5.9}$$

hence $W = A^{\top}Q\Lambda^{-1}Q^{\top}$, as required.

The derivation of an explicit form for the matrix W, mapping gapped k-mers to ℓ-mers, is the central result of this chapter. Although equation (5.3) gives a method to obtain the weight matrix W from the eigendecomposition of matrix $AA^\top$ given in 5.6, numerical calculation of the eigenvectors for $AA^\top$ is only feasible for small values of ℓ, k, and b, as the size of matrix A grows rapidly with larger values of ℓ, k, and b. For example, for $(\ell = 15, k = 7, b = 4)$, we have $N \cong 10^9$, and $M \cong 10^8$. However, considering the symmetry in matrix A, it is easily proved that the matrix W has a simple structure: the entry $w_{i,j}$ only depends on the number of mismatches between the ℓ-mer u_i and the gapped-kmer v_j; in other words, there exists a finite sequence of values $w_0, w_1, \cdots, w_k$ such that $w_{i,j} = w_m$ if u_i and v_j have exactly m mismatches. Thus, the entries of matrix W are limited to a small set of values $\{w_0, \ldots, w_k\}$. These values are specified by the following theorem:

Theorem 5.2.2 *The values of the elements of matrix W are given by the following equation, in which, ℓ is the sequence length, b is the size of the alphabet, k is the number of known bits, and m is the number of mismatches between the corresponding ℓ-mer u_i and the gapped-kmer v_j:*

$$w_{\ell,k,m} = \frac{\binom{k-\ell}{m}}{b^\ell \binom{\ell}{k}\binom{k}{m}} \sum_{n=0}^{k-m} \binom{\ell}{n} (b-1)^n \tag{5.10}$$

We will prove this by first finding an eigendecomposition for matrix $AA^\top$ and then applying (5.3) in the following sections. First we will give some definitions.

5.3 Preliminaries and Notation

In this section, first we review some linear algebra definitions and theorems in (5.3.1), then provide some binomial identities in (5.3.2), then define some alphabets and word sets and matrix A in (5.3.3). Finally in subsection (5.3.4) we define function ν and prove some identities that enable us to find an eigendecomposition for matrix $AA^{\top}$.

5.3.1 Review of Some Linear Algebra Theorems

Let F be a field and $A \in F^{m\times n}$ be a matrix. The row space of A is denoted as $\text{row}(A)$, the column space of A is denoted as $\text{col}(A)$, and the the dimension of the row space (which is the same as the dimension of the column space) of A is denoted as $\text{rank}(A)$. The kernel of A, denoted as $\ker(A)$, is the space of all column vectors x satisfying $Ax = \mathbf{0}$ and the dimension of this space is called the nullity of A and denoted as $\text{null}(A)$. It is known that $\text{null}(A)+\text{rank}(A) = n$. Let $B \in F^{n\times n}$. The characteristic polynomial of B is defined as $p_B(z) = \det(zI-X)$. An element $\lambda \in F$ is an eigenvalue of B if there exists a nonzero column vector x satisfying $Bx = \lambda x$; The vector x is called an eigenvector of B. It is observed that λ is an eigenvalue of B if and only if it is a root of the characteristic equation $p_B(z)$. For an eigenvalue λ, the space $\ker(B-\lambda I)$ is called the eigenspace of B corresponding to λ. The *algebraic multiplicity* of an eigenvalue λ, denoted as $\alpha(\lambda)$ is the multiplicity of the root λ of $p_B(z)$. The geometric

multiplicity of an eigenvalue λ, denoted by $\gamma(\lambda)$, is the dimension of $\ker(B-\lambda I)$. The matrix B is called diagonalizable if there exists a nonsingular matrix $P \in F^{n\times n}$ such that $B = P\Lambda_0 P^{-1}$ for some diagonal matrix $\Lambda_0 \in F^{n\times n}$. A matrix $B \in \mathcal{R}^{n\times n}$ is known to be diagonalizable by a nonsingular matrix P if and only if the eigenvalues of B are real and $\alpha(\lambda) = \gamma(\lambda)$ for all eigenvalues λ of B. It is deduced that a diagonalizable matrix B has a diagonal decomposition $B = P\Lambda_0 P^\top$ where P is a unitary matrix whose columns contain normalized orthogonal eigenvectors of B and Λ_0 is a diagonal matrix having eigenvalues of B on its diagonal, ordered according to the columns of P. From $B = P\Lambda_0 P^\top$ by simple matrix calculations, we conclude that $B = Q\Lambda Q^\top$, where the matrix Q is obtained by deleting the columns of P which are in $\ker(B)$, and Λ) is obtained by deleting the zero columns of Λ_0.

For a matrix $A \in \mathcal{C}^{m\times n}$ its Hermitian adjoint, A^*, is its conjugate transpose, i.e. A^* is an $n \times m$ matrix with $A^*(i,j) = \overline{A(j,i)}$. A matrix A is Hermitian if $A = A^*$, thus a real matrix A is Hermitian if and only if it is symmetric. A Hermitian matrix A is positive definite if $x^*Ax > 0$ for all nonzero $x \in \mathcal{C}^n$, and positive semi-definite ($x^*Ax \geq 0$) for all nonzero $x \in \mathcal{C}^n$. It is concluded that a matrix $A \in S_n$ is positive definite if $x^\top Ax > 0$ for all nonzero $x \in \mathcal{R}^n$, and positive semi-definite if $x^\top Ax \geq 0$ for all nonzero $x \in \mathcal{R}^n$.

For any matrix A, the matrix $A^\top A$ is positive semidefinite, and $\text{rank}(A) = \text{rank}(AA^\top)$. Conversely, any Hermitian positive semidefinite matrix M can be written as $M = A^*A$; this is the Cholesky decomposition.

The Moore-Penrose pseudoinverse of a matrix A, denoted by A^+, is defined as a matrix that satisfies all the following four conditions:

$$AA^+A = A$$

$$A^+AA^+ = A^+$$

$$(AA^+)^* = AA^+$$

$$(A^+A)^* = A^+A.$$

The Moore-Penrose pseudoinverse exists and is unique for any given matrix A. Also we have $A^+ = (A^*A)^+A^* = A^*(AA^*)^+$. For further properties of the Moore-Penrose pseudoiverse see for instance [100].

5.3.2 Some Binomial Identities

For real number x and nonnegative integer n, the falling factorial $(x)_n$ is defined as $x(x-1)\cdots(x-n+1)$ and the binomial coefficient $\binom{x}{n}$ is defined as $\frac{(x)_n}{n!}$. The binomial coefficients satisfy many interesting identities, some of which are discussed in this section (for further information see [101]). The following examples are among the well-known identities used frequently in this chapter:

$$\binom{r}{s}\binom{s}{t} = \binom{r}{t}\binom{r-t}{s-t},$$

$$\sum_{i=0}^{k}\binom{i+n}{i} = \binom{k+n+1}{k},$$

$$\sum_{i=0}^{k}(-1)^i\binom{n}{i} = \binom{k-n}{k}.$$

Two more identities are given in the following Proposition.

Proposition 5.3.1 (i) *Let $0 \le t \le n \le p \le k$ be integers. Then the following identity holds:*

$$\binom{k}{p}\binom{k-p}{n-t}\binom{p}{t} = \binom{k}{n}\binom{n}{t}\binom{k-n}{p-t} \tag{5.11}$$

(ii) *Let $0 \le p \le k \le \ell$ be integers. Then the following identity holds:*

$$\sum_{n=0}^{k}\sum_{t=0}^{n}(-1)^{n-t}\binom{\ell}{n}\binom{n}{t}\binom{k-n}{p-t}x^t = \binom{k-\ell}{k-p}\sum_{n=0}^{p}\binom{\ell}{n}x^n. \tag{5.12}$$

Proof The first identity is proved as follows:

$$\binom{k}{p}\binom{p}{t}\binom{k-p}{n-t} = \binom{k}{t}\binom{k-t}{p-t}\binom{k-p}{n-t} = \binom{k}{t}\binom{k-t}{p-t}\binom{k-p}{k-n-p+t} =$$
$$\binom{k}{t}\binom{k-t}{k-n}\binom{k-n}{p-t} = \binom{k}{t}\binom{k-t}{n-t}\binom{k-n}{p-t} = \binom{k}{n}\binom{n}{t}\binom{k-n}{p-t}.$$

To prove (5.12), denote the left side by $\tau_{p,k,\ell}$, then using $\binom{k-n}{p-t} = \binom{k-n-1}{p-t-1} + \binom{k-n-1}{p-t}$, we obtain

$$\begin{aligned}
\tau_{p,k,\ell} &= \tau_{p-1,k-1,\ell} + \sum_{t=0}^{k}(-1)^{k-t}\binom{\ell}{k}\binom{k}{t}\binom{-1}{p-t-1}x^t + \tau_{p,k-1,\ell} + \sum_{t=0}^{k}(-1)^{k-t}\binom{\ell}{k}\binom{k}{t}\binom{-1}{p-t}x^t \\
&= \tau_{p-1,k-1,\ell} + \tau_{p,k-1,\ell} + \sum_{t=0}^{k}(-1)^{k-t}\binom{\ell}{k}\binom{k}{t}\left(\binom{-1}{p-t} + \binom{-1}{p-t-1}\right)x^t \\
&= \tau_{p-1,k-1,\ell} + \tau_{p,k-1,\ell} + \sum_{t=0}^{k}(-1)^{k-t}\binom{\ell}{k}\binom{k}{t}\binom{0}{p-t}x^t \\
&= \tau_{p-1,k-1,\ell} + \tau_{p,k-1,\ell} + (-1)^{k-p}\binom{\ell}{k}\binom{k}{p}x^p.
\end{aligned}$$

Then by replacing k by k' in $\tau_{p,k,\ell} - \tau_{p,k-1,\ell} = (-1)^{k-p}\binom{\ell}{p}\binom{\ell-p}{k-p}x^p + \tau_{p-1,k-1,\ell}$ and summing up over k', $0 \le k' \le k$, we get

$$\tau_{p,k,\ell} = \sum_{k'=0}^{k}(-1)^{k'-p}\binom{\ell}{p}\binom{\ell-p}{k'-p}x^p + \sum_{k'=0}^{k}\tau_{p-1,k'-1,\ell},$$

whence

$$\tau_{p,k,\ell} = \binom{\ell}{p}\binom{k-\ell}{k-p}x^p + \sum_{k'=0}^{k} \tau_{p-1,k'-1,\ell}. \tag{5.13}$$

Now we prove the identity $\tau_{p,k,\ell} = \binom{k-\ell}{k-p}\sum_{n=0}^{p}\binom{\ell}{n}x^n$ for $p = 0, 1, \cdots, k$ by bounded induction on p. For $p = 0$, using (5.13) we obtain $\tau_{0,k,\ell} = \binom{k-\ell}{k}$ which proves the required identity in this case. Now let $0 < p \leq k$ and suppose that the result is true for $p-1$. We prove the validity of the identity for p by using the induction hypothesis and (5.13), as follows:

$$\begin{aligned}
\tau_{p,k,\ell} &= \binom{k-\ell}{k-p}\binom{\ell}{p}x^p + \sum_{k'=0}^{k} \tau_{p-1,k'-1,\ell} \\
&= \binom{k-\ell}{k-p}\binom{\ell}{p}x^p + \sum_{k'=0}^{k}\binom{k'-\ell-1}{k'-p}\sum_{n=0}^{p-1}\binom{\ell}{n}x^n \\
&= \binom{k-\ell}{k-p}\binom{\ell}{p}x^p + \binom{k-\ell}{k-p}\sum_{n=0}^{p-1}\binom{\ell}{n}x^n \\
&= \binom{k-\ell}{k-p}\sum_{n=0}^{p}\binom{\ell}{n}x^n.
\end{aligned}$$

5.3.3 Some More Definitions and Notations

Since we work with sequences of given symbols (or words over on a given alphabet), some definitions in this respect are mentioned. Let $b > 1$ be an integer; as usual, we define the b-ary alphabet by $\Sigma_b = \{0, 1, \cdots, b-1\}$. In addition to this alphabet, two other alphabets, defined as follows, are frequently used in this work:

$$\Delta_b = \Sigma_b \cup \{g\}$$

$$\Gamma_b = \Delta_b \setminus \{0\}$$

The symbol g used above represents "gap" (or wildcard) symbol. Hence, a gapped k-mer v of length ℓ is a sequence v of length ℓ over $\Delta_b \cup \{g\}$ with exactly $\ell - k$ occurrences of symbol g. We add some useful definitions from the literature to simplify the representation of our problem here. A word w on a given alphabet $\mathcal{A}$ is a finite sequence $w = w_1 w_2 \cdots w_n$ with $w_i \in \mathcal{A}$. The length of w, the integer n, is denoted as $|w|$. The occurrences of a symbol i in a word w is denoted as $|w|_i$. Five important sets of words are defined below:

$$U_\ell = \Sigma_b^\ell$$

$$V_{\ell k} = \{w \in \Delta_b^\ell : |w|_g = \ell - k\}$$

$$V'_{\ell k} = \{w \in \Gamma_b^\ell : |w|_g = \ell - k\}$$

$$V_{\ell,\leq k} = \bigcup_{m=0}^{k} V_{\ell m}$$

$$V'_{\ell,\leq k} = \bigcup_{m=0}^{k} V'_{\ell m}$$

For a set X and a nonnegative integer n, by $\binom{X}{n}$, we mean the set of all n-element subsets of X. For a word $v' \in \Gamma_b^*$, we let $G_{v'} = \{i : 1 \leq i \leq |v'|, v'_i = g\}$ and $\overline{G}_{v'} = \{1, \cdots, |v'|\} \setminus G_{v'}$. Thus if $v' \in \Gamma_b^\ell$, then we have $|\overline{G}_{v'}| = n$ if and only if $v' \in V'_{\ell n}$. We say elements $u \in U_\ell$ and $v \in V_{\ell k}$ match if for any $1 \leq i \leq \ell$ with $v_i \neq g$ we have $u_i = v_i$. The set of the elements $v \in V_{\ell k}$ which are matchable with u is denoted by $M_{\ell k}(u)$. The set of elements $u \in U_\ell$ which are matchable with v is denoted by $M'_{\ell k}(v)$.

We define the matrix $A_{\ell k}$ as a $(0,1)$-matrix whose rows (columns) are indexed

by elements of $V_{\ell k}$ (elements of U_ℓ) and with $A_{\ell k}(v,u) = 1$ if and only if u and v match. Thus we have $A_{\ell k}(v,u) = 1$ if and only if $v \in M_{\ell k}(u)$ if and only if $u \in M'_{\ell k}(v)$. If we interpret each element $v \in V_{\ell k}$ as the set of all elements $u \in U_\ell$ which are matchable with v, then an incidence structure with the point-set U_ℓ and the block-set $V_{\ell k}$ is obtained in which every block consists of $b^{\ell-k}$ points and every point occurs in $\binom{\ell}{k}$ blocks. The matrix $A_{\ell k}^\top$ then appears as the incidence matrix of this incidence structure which is simultaneously the inclusion matrix of elements of U_ℓ vs. the elements of $V_{\ell k}$.

In this chapter we prove that $A_{\ell k}A_{\ell k}^\top$ is diagonalizable and based on an explicit eigendecomposition we find the Moore-Penrose pseudo-inverse of $A_{\ell k}$ and we provide a formula for the entries of this matrix.

5.3.4 Function ν and Related Identities

For many applications it is useful to study the matrix $A_{\ell i}A_{\ell j}^\top$. The rows and the columns of this matrix are respectively indexed by elements of $V_{\ell k}$ and the entry $\left(A_{\ell k}A_{\ell k}^\top\right)(v,w)$ has a simple combinatorial interpretation: The value $\left(A_{\ell k}A_{\ell k}^\top\right)(v,w)$ counts the number of elements of U_ℓ which match both v and w. In this section we present an orthogonal basis for the eigenspace of $A_{\ell k}A_{\ell k}^\top$. For this purpose, first we introduce the definition of the function $\nu : \Delta_b \times \Delta_b \to Z$ which has an essential role in the rest of this section. Then by using some identities we find the eigenvalues and the corresponding eigenvectors. We extend the natural order on (nonnegative) integers

to $Z \cup \{g\}$ using the convention $i < g$ for any integer i. Then we define the function $\nu : \Delta_b \times \Delta_b \to Z$ as follows:

$$\nu(x,y) = \begin{cases} 1 & \text{if } x < y \text{ or } x = y = g, \\ 0 & \text{if } x > y, \\ -y & \text{if } x = y \neq g. \end{cases}$$

For any positive integer ℓ, this definition is naturally extended to $\nu : \Delta_b^\ell \times \Delta_b^\ell \to Z$ by a simple product rule as follows:

$$\nu(w_1 \cdots w_\ell, v'_1 \cdots v'_\ell) = \prod_{i=1}^{\ell} \nu(w_i, v'_i).$$

Let n, k and ℓ be integers with $0 \leq n \leq k \leq \ell$. Recall that $V'_{\ell n}$ is the set of all words over the alphabet Γ_b with exactly $\ell - n$ symbols g. Below we define two column vectors which are very important in this section.

Definition The vector $x_{v'}^{\ell k n}$ is a column vector whose rows are indexed by the elements of $V_{\ell k}$ with entries $x_{v'}^{\ell k n}(w) = \nu(w, v')$. The column vector $z_{v'}^{\ell n}$ is defined by $z_{v'}^{\ell n} = x_{v'}^{\ell \ell n}$; In other words, $z_{v'}^{\ell n}$ is a column vector whose rows are indexed by elements u of U_ℓ with entries $z_{v'}^{\ell k}(u) = \nu(u, v')$. When there is no need to emphasize the parameters ℓ, k and n, we simply write $x_{v'}$ and $z_{v'}$.

Proposition 5.3.2 *Let $n \leq k \leq \ell$ and $v' \in V'_{\ell n}$. The following matrix identities hold.*

(i) $A_{\ell k}^\top x_{v'}^{\ell k n} = \binom{\ell - n}{\ell - k} z_{v'}^{\ell n}$.

(ii) $A_{\ell k} z_{v'}^{\ell n} = b^{\ell-k} x_{v'}^{\ell k n}$.

(iii) $A_{\ell i} A_{\ell j}^{\top} x_{v'}^{\ell j n} = b^{\ell-i} \binom{\ell-n}{\ell-j} x_{v'}^{\ell i n}$.

(iv) $A_{\ell k}^{\top} A_{\ell k} z_{v'}^{\ell n} = b^{\ell-k} \binom{\ell-n}{\ell-k} z_{v'}^{\ell k}$.

Proof (i) It is enough to prove that the identity

$$\sum_{y \in M_{\ell k}(u)} \nu(y, v') = \binom{\ell - n}{\ell - k} \nu(u, v') \tag{5.14}$$

holds for any $u \in U_\ell$. The nonzero summands of the summation on the left are obtained from the words $y \in V_{\ell k}$ in which all symbols g appear in positions $G_{v'}$ (Using $|G_{v'}| = \ell - n$, we conclude that there are $\binom{\ell-n}{\ell-k}$ such summands). On the other hand, since $y \in M_{\ell k}(u)$, it is easily seen that for any such y we have $\nu(y, v') = \nu(u, v')$. Thus the summation is simplified to $\binom{\ell-n}{\ell-k}\nu(u, v')$ as required.

(ii) It is enough to prove that the identity

$$\sum_{u \in M'_{\ell k}(v)} \nu(u, v') = b^{\ell-k} \nu(v, v') \tag{5.15}$$

holds for any $v \in V_{\ell k}$. We prove this in two cases:

Case (a). Suppose that $G_v \subseteq G_{v'}$. Then for any $u \in M'_{\ell k}(v)$ we have

$$\nu(u, v') = \prod_{i \in \overline{G}_{v'}} \nu(u_i, v'_i) = \prod_{i \in \overline{G}_{v'}} \nu(v_i, v'_i) = \nu(v, v')$$

and since there are $b^{\ell-k}$ such words u, the equation (5.15) thus follows.

Case (b). Suppose that $G_v \not\subseteq G_{v'}$, consequently $G_v \setminus G_{v'} \neq \emptyset$. Now for any $i \in G_v \setminus G_{v'}$ we have $\nu(v_i, v_i') = 0$, thus the right side of (5.15) is 0. We prove that the left side is also 0 as follows. The nonzero summands in the summation are obtained from elements $u \in X$ where the subset $X \subseteq U_\ell$ is given by

$$X = \{u \in U_\ell : u_i \leq v_i \text{ for } i \in G_v \setminus G_{v'} \text{ and } u_i = v_i \text{ for } i \in \overline{G}_v\}.$$

Thus we obtain

$$\begin{aligned}
\sum_{u \in M'_{\ell k}(v)} \nu(u, v') &= \sum_{u \in X} \nu(u, v') \\
&= \sum_{u \in X} \prod_{i=1}^{\ell} \nu(u_i, v_i') \\
&= \sum_{u \in X} \left(\prod_{i \in \overline{G}_v} \nu(u_i, v_i') \prod_{i \in G_v \setminus G_{v'}} \nu(u_i, v_i') \prod_{i \in G_v \cap G_{v'}} \nu(u_i, v_i') \right) \\
&= \sum_{u \in X} \left(\prod_{i \in \overline{G}_v} \nu(v_i, v_i') \prod_{i \in G_v \setminus G_{v'}} \nu(u_i, v_i') \prod_{i \in G_v \cap G_{v'}} \nu(u_i, g) \right) \\
&= \prod_{i \in \overline{G}_v} \nu(v_i, v_i') \prod_{i \in G_v \setminus G_{v'}} \sum_{u_i=0}^{v_i'} \nu(u_i, v_i') \prod_{i \in G_v \cap G_{v'}} \sum_{u_i=0}^{b-1} \nu(u_i, g) \\
&= b^{|G_v \cap G_{v'}|} \prod_{i \in \overline{G}_v} \nu(v_i, v_i') \prod_{i \in G_v \setminus G_{v'}} \sum_{u_i=0}^{v_i'} \nu(u_i, v_i') \\
&= 0.
\end{aligned}$$

The lThe last identity holds because for any $i \in G_v \setminus G_{v'}$ we have $\sum_{u_i=0}^{v_i'} \nu(u_i, v_i') = \sum_{u_i=}^{v_i'-} \sum_{u_i=0}^{v_i'-1} 1 - v_i' = 0$.

(iii),(iv) These are immediate consequences of parts (i) and (ii).

Remark 1. The definition of the function ν and its extension and the inverse type formulas of parts parts (i) and (ii) of the previous proposition show similarities with properties of the Möbius function (see Chapter 12 of [87] for instance).

Proposition 5.3.3 *Let $u \in U_\ell$ and $v \in V_{\ell k}$. Let $P = \{i : 1 \leq i \leq \ell, v_i \neq g, v_i = u_i\}$, $Q = \{i : 1 \leq i \leq \ell, v_i \neq g, v_i \neq u_i\}$, $|P| = p$ (and consequently $|Q| = k - p$).*

(i) *For $1 \leq i \leq \ell$ let*

$$\phi_i(u, v) = \sum_{j=\max\{1,v_i\}}^{b-1} \frac{\nu(v_i, j)\nu(u_i, j)}{j(j+1)}.$$

Then we have

$$\phi_i(u, v) = \begin{cases} \frac{b-1}{b}, & \text{if } i \in P; \\ \\ \frac{-1}{b}, & \text{otherwise, i.e. if } i \in Q. \end{cases}$$

(ii) *The following identity holds*

$$\sum_{v' \in V'_{\ell n}} \frac{\nu(v, v')\nu(u, v')}{\prod_{i \in \overline{G}_{v'}} v'_i(v'_i + 1)} = \frac{1}{b^n} \sum_{t=0}^{n} (-1)^{n-t} \binom{p}{t} \binom{k-p}{n-t} (b-1)^t, \tag{5.16}$$

Proof (i) If $i \in P$ and $v_i > 0$ then

$$\begin{aligned}
\phi_i(u,v) &= \sum_{j=v_i}^{b-1} \frac{\nu(v_i,j)\nu(u_i,j)}{j(j+1)} \\
&= \sum_{j=v_i}^{b-1} \frac{\nu(v_i,j)^2}{j(j+1)} \\
&= \frac{v_i^2}{v_i(v_i+1)} + \sum_{j=v_i+1}^{b-1} \frac{1}{j(j+1)} \\
&= \frac{v_i}{v_i+1} + \left(\frac{1}{v_i+1} - \frac{1}{b}\right) \\
&= \frac{b-1}{b}.
\end{aligned}$$

If $i \in P$ and $v_i = 0$, then given $j \geq 1$, we have $\nu(u_i,j) = \nu(v_i,j) = 1$, hence,

$$\begin{aligned}
\phi_i(u,v) &= \sum_{j=1}^{b-1} \frac{1}{j(j+1)} \\
&= \frac{b-1}{b}.
\end{aligned}$$

The case $i \in Q$ is done similarly.

(ii) Without loss of generality suppose that $v = v_1 \cdots v_k g^{\ell-k}$. Let $P = \{1, \cdots, p\}$ and $Q = \{p+1, \cdots, k\}$ and for a subset $\overline{G} \in \binom{\{1,\cdots,\ell\}}{\ell-n}$ let $X'_{\ell n}(\overline{G}) = \{v' \in V'_{\ell n} : \overline{G}_{v'} = \overline{G}\}$. Denote the left side of (5.16) by S. Since the nonzero summands in (5.16) are obtained from the words $v' \in V'_{\ell n}$ which have g's at the $\ell - k$ rightmost positions, we may just consider words $v' = v'_1 \cdots v'_k g^{\ell-k}$ with $|v'_1 \cdots v'_k|_g = k - n$.

Then we have

$$
\begin{aligned}
S &= \sum_{t=0}^{n} \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \sum_{v' \in X'_{\ell n}(T_1 \cup T_2)} \frac{\nu(v,v')\nu(u,v')}{\prod_{i \in \overline{G}_{v'}} v'_i(v'_i+1)} \\
&= \sum_{t=0}^{n} \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \sum_{v' \in X'_{\ell n}(T_1 \cup T_2)} \frac{\prod_{i \in \overline{G}_{v'}} \nu(v_i,v'_i)\nu(u_i,v'_i)}{\prod_{i \in \overline{G}_{v'}} v'_i(v'_i+1)} \\
&= \sum_{t=0}^{n} \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \sum_{v' \in X'_{\ell n}(T_1 \cup T_2)} \prod_{i \in T_1 \cup T_2} \frac{\nu(v_i,v'_i)\nu(u_i,v'_i)}{v'_i(v'_i+1)} \\
&= \sum_{t=0}^{n} \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \prod_{i \in T_1 \cup T_2} \left(\sum_{v'_i = \max\{1, v_i\}}^{b-1} \frac{\nu(v_i,v'_i)\nu(u_i,v'_i)}{v'_i(v'_i+1)} \right) \\
&= \sum_{t=0}^{n} \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} \prod_{i \in T_1 \cup T_2} \phi_i(u,v) \\
&= \sum_{t=0}^{n} \sum_{T_1 \in \binom{P}{t}} \sum_{T_2 \in \binom{Q}{n-t}} (\frac{b-1}{b})^t (\frac{-1}{b})^{n-t} \quad \text{(by part (i))} \\
&= \frac{1}{b^n} \sum_{t=0}^{n} (-1)^{n-t} \binom{p}{t} \binom{k-p}{n-t} (b-1)^t.
\end{aligned}
$$

5.4 The Eigendecomposition of $A_{\ell k} A_{\ell k}^{\top}$

Definition Given b, we define $\Delta_{\ell k}$ as a matrix whose rows and columns are indexed by the elements of $V_{\ell k}$ and $V'_{\ell,\leq k}$ respectively and whose entries are given by $\Delta_{\ell k}(w, v') = \nu(w, v')$. In other words, the columns of $\Delta_{\ell k}$ are exactly the vectors $x_{v'}$ for $v' \in V'_{\ell,\leq k}$. For $0 \leq m \leq k$, we define the matrix $\Delta_{\ell,k;m}$ as the submatrix of $\Delta_{\ell,k}$ which is obtained by retaining the columns indexed by elements $v' \in V'_{\ell m}$ and omitting

the other columns. Consequently, we have the following block decomposition:

$$\Delta_{\ell,k} = [\Delta_{\ell,k;0}, \Delta_{\ell,k;1}, \cdots, \Delta_{\ell,k;k}]$$

The block decomposition of $\Delta_{\ell,k;m}$ is given in the proof of the following proposition.

Proposition 5.4.1 *The matrix $A_{\ell k}A_{\ell k}^{\top}$ has exactly $\sum_{n=0}^{k}\binom{\ell}{n}(b-1)^n$ nonzero eigenvalues as follows: For any n, $0 \leq n \leq k$, there are $\binom{\ell}{n}(b-1)^n$ eigenvalues equal to $\lambda_n = \binom{\ell-n}{k-n}b^{\ell-k}$ and the eigenvectors corresponding λ_n are of the form $x_{v'}^{\ell kn}$ with $v' \in V'_{\ell n}$. These eigenvectors are pairwise orthogonal and $A_{\ell k}A_{\ell k}^{\top}$ is diagonalizable.*

Proof By using Proposition 5.3.2 (iii), each $x_{v'}^{\ell kn}$ is an eigenvector of $A_{\ell k}A_{\ell k}^{\top}$ corresponding to the eigenvalue $\binom{\ell-n}{k-n}b^{\ell-k}$. Moreover, for different words $u' \in V'_{\ell,n_1}$ and $v' \in V'_{\ell,n_2}$, the vectors $x_{u'}^{\ell,k,n_1}$ and $x_{v'}^{\ell,k,n_2}$ are orthogonal as it is proved in the sequel. To prove the orthogonality of the eigenvectors, it is enough to prove that for any integer ℓ, for all integers k, m, n with $0 \leq m, n \leq k$ the matrix products $\Delta_{\ell k;n}^{\top}\Delta_{\ell k;m}$ are diagonal if $n = m$ and $\mathbf{0}$ otherwise. We prove this by induction on ℓ. For $\ell = 0$ there exists just one matrix $\Delta_{0,k;n}$, so there is nothing to prove. For $\ell = 1$, either $k = 0$ or $k = 1$. In the first case the matrix $\Delta_{1,0;0}$ has just one entry, so there is nothing to prove. In the case $k = 1$ we have the following matrices and it is straightforward to check that the matrices $\Delta_{1,1;n}^{\top}\Delta_{1,1;m}$ are diagonal if $n = m$ and $\mathbf{0}$ otherwise.

$$\Delta_{1,1;0} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ b-1 \end{array} \begin{array}{c} g \\ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ \vdots \end{pmatrix} \end{array}$$

$$\Delta_{1,1;1} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ i-1 \\ i \\ \vdots \\ \vdots \\ \vdots \\ b-1 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & \cdots & i & \cdots & \cdots & b-1 \end{array} \\ \begin{pmatrix} 1 & 1 & \cdots & 1 & \cdots & \cdots & 1 \\ -1 & 1 & \ddots & 1 & \ddots & \ddots & 1 \\ 0 & -2 & \ddots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \cdots & 0 & -(b-1) \end{pmatrix} \end{array}$$

Before continuing the proof, note that from the definition of $\Delta_{\ell,k;m}$ the block decomposition of this matrix is concluded as follows, where $0\cdots$, for instance, stands for the set of all indexes commencing with 0. Note that the matrix $\mathbf{0}_{\ell k;n}$ is an $\binom{\ell}{k}b^k \times \binom{\ell}{n}(b-1)^n$ matrix with all zero entries.

$$
\begin{array}{c|c|c|c|c|c|c}
 & g\cdots & 1\cdots & \cdots & i\cdots & \cdots & b-1\cdots \\
\hline
0\cdots & \Delta_{\ell-1,k-1;m} & \Delta_{\ell-1,k-1;m-1} & \cdots & \Delta_{\ell-1,k-1;m-1} & \cdots & \Delta_{\ell-1,k-1;m-1} \\
1\cdots & \Delta_{\ell-1,k-1;m} & -\Delta_{\ell-1,k-1;m-1} & \ddots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \mathbf{0}_{\ell-1,k-1;m-1} & \ddots & \Delta_{\ell-1,k-1;m-1} & \ddots & \vdots \\
i\cdots & \vdots & \vdots & \ddots & -i\Delta_{\ell-1,k-1;m-1} & \ddots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \mathbf{0}_{\ell-1,k-1;m-1} & \ddots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \Delta_{\ell-1,k-1;m-1} \\
b-1\cdots & \Delta_{\ell-1,k-1;m} & \mathbf{0}_{\ell-1,k-1;m-1} & \cdots & \mathbf{0}_{\ell-1,k-1;m-1} & \cdots & \mathbf{0}_{\ell-1,k-1;m-1} \quad -(b-1)\Delta_{\ell-1,k-1;m-1} \\
\hline
g\cdots & \Delta_{\ell-1,k;m} & \mathbf{0}_{\ell-1,k;m-1} & \cdots & \mathbf{0}_{\ell-1,k;m-1} & \cdots & \cdots \quad \mathbf{0}_{\ell-1,k;m-1}
\end{array}
$$

Now we suppose that the aforementioned property holds for $\ell-1$, which means that the matrices $\Delta_{\ell-1,k;n}^{\top}\Delta_{\ell-1,k;m}$ are diagonal if $n = m$ and $\mathbf{0}$ otherwise. To prove this property for ℓ, note that using the above block decomposition and some calculation, the matrix $\Delta_{\ell,k;n}^{\top}\Delta_{\ell,k;m}$ consists of $b \times b$ blocks, all of which are $\mathbf{0}$ when $m \neq n$; but the matrices on the diagonal are themselves diagonal when $m = n$. Considering the dimensions, this fact simply implies the required proposition for ℓ and proves the result.

Now note that the matrix $A_{\ell k}A_{\ell k}^{\top}$ has entries $b^{\ell-k}$ on its main diagonal, therefore $trace(A_{\ell k}A_{\ell k}^{\top}) = b^{\ell-k}b^k\binom{\ell}{k} = b^{\ell}\binom{\ell}{k}$ and this equals the summation of previously obtained eigenvalues, considering their multiplicities, as follows:

$$\sum_{n=0}^{k}\binom{\ell}{n}(b-1)^n\lambda_n = \sum_{n=0}^{k}\binom{\ell}{n}\binom{\ell-n}{k-n}b^{\ell-k}(b-1)^n = b^{\ell-k}\binom{\ell}{k}\sum_{n=0}^{k}\binom{k}{n}(b-1)^n = b^{\ell}\binom{\ell}{k}.$$

But the matrix $A_{\ell k}A_{\ell k}^{\top}$ is symmetric positive semi-definite, so it has no negative eigenvalues; hence, all the remaining eigenvalues are 0. We have already shown that

for any nonzero eigenvalue λ of $A_{\ell k}A_{\ell k}^{\top}$, the algebraic multiplicity and the geometric multiplicity of λ are the same. For $\lambda = 0$, observe that $A_{\ell k}A_{\ell k}^{\top}$ is a square matrix of order $\binom{\ell}{k}b^k$ and the algebraic multiplicity of 0 is $\alpha(0) = \binom{\ell}{k}b^k - \sum_{n=0}^{k}\binom{\ell}{n}(b-1)^n$. On the other hand, the geometric multiplicity of 0 is the nullity of $A_{\ell k}A_{\ell k}^{\top}$ which equals $\gamma(0) = \binom{\ell}{k}b^k - \text{rank}(A_{\ell k}A_{\ell k}^{\top})$, hence $\gamma(0) = \alpha(0)$. Since for any eigenvalue λ of $A_{\ell k}A_{\ell k}^{\top}$, the equation $\gamma(\lambda) = \alpha(\lambda)$ holds, we conclude that this matrix is diagonalizable.

Remark 2. Again consider equation (5.3) and suppose that the matrix $A_{\ell k}A_{\ell k}^{\top}$ has an eigendecomposition of the form $A_{\ell k}A_{\ell k}^{\top} = P_{\ell k}\Lambda_0 P_{\ell k}^{\top}$ in which the columns of $P_{\ell k}$ are normalized orthogonal eigenvectors and the diagonal matrix Λ_0 has corresponding eigenvalues on its diagonal. Moreover, we may suppose that the nonzero eigenvalues appear first, thus the matrix $P_{\ell k}$ has a decomposition $P_{\ell k} = [Q_{\ell k}\ N_{\ell k}]$ where the columns of $Q_{\ell k}$ are the eigenvectors which do not lie in the null space of $A_{\ell k}A_{\ell k}^{\top}$, and the columns of $N_{\ell k}$ are the eigenvectors which lie in the null space of $A_{\ell k}A_{\ell k}^{\top}$.

More precisely, we present the matrix $Q_{\ell k}$ as below.

Definition Based on the informal description of $Q_{\ell k}$ given in Remark 2 and the results of Proposition 5.4.1, we define $Q_{\ell k}$ as a matrix whose columns are vectors $\frac{1}{||x_{v'}||}x_{v'}$, where $v' \in V'_{\ell,\leq k}$.

Proposition 5.4.2 *With the above definitions, the matrix $A_{\ell k}A_{\ell k}^{\top}$ admits an eigendecomposition of the form $A_{\ell k}A_{\ell k}^{\top} = Q_{\ell k}\Lambda^{-1}Q_{\ell k}^{\top}$. Moreover, the following properties hold:*

(i)

$$Q_{\ell k} = \Delta_{\ell k} E, \tag{5.17}$$

where $E = \mathrm{diag}(\frac{1}{||x_{v'}||})_{v' \in V'_{\ell, \leq k}}$. *Moreover for* $v' \in V'_{\ell n}$ *we have*

$$||x_{v'}^{\ell k n}|| = \sqrt{\binom{\ell - n}{\ell - k} b^{k-n} \prod_{i \in \overline{G}_{v'}} (v'_i + v'^2_i)}. \tag{5.18}$$

(ii) *We have*

$$Q_{\ell k} \Lambda Q_{\ell k}^{\top} = A_{\ell k} A_{\ell k}^{\top}, \tag{5.19}$$

$$Q_{\ell k} \Lambda^{-1} Q_{\ell k}^{\top} = \Delta_{\ell k} D \Delta_{\ell k}^{\top}, \tag{5.20}$$

where $D = diag\left((d_{v'})_{v' \in V'_{\ell, \leq k}}\right)$ *and*

$$d_{v'} = \frac{1}{\binom{\ell - |\overline{G}_{v'}|}{\ell - k}^2 b^{\ell - |\overline{G}_{v'}|} \prod_{i \in \overline{G}_{v'}} (v'_i + v'^2_i)}. \tag{5.21}$$

(iii) *We have*

$$Q_{\ell k}^{\top} Q_{\ell k} = I, \tag{5.22}$$

$$Q_{\ell k} Q_{\ell k}^{\top} A_{\ell k} = A_{\ell k}. \tag{5.23}$$

Proof (i) These are immediate consequences of Definition 3, except the last equation. To calculate the norm of $x_{v'}$, without loss of generality let $\overline{G}_{v'} = \{1, \cdots n\}$, hence $v' = v'_1 \cdots v'_n g^{\ell - n}$ and it is evident that the nonzero elements $\nu(w, v')$ of $x_{v'}$ are obtained from the words $w = rs \in V_{\ell k}$ with $|s| = \ell - n$, $|s|_g = \ell - k$,

$r = r_1 \cdots r_n$ and $0 \leq r_i \leq v'_i$ for $i = 1, \cdots, n$. Moreover, the value $\nu(w, v')$ is independent of s and there are $\binom{\ell-n}{\ell-k} b^{k-n}$ choices for s. We thus provide

$$\begin{aligned} ||x_{v'}||^2 &= \binom{\ell-n}{\ell-k} b^{k-n} \prod_{i=1}^{n} \sum_{x_i=0}^{v'_i} \nu(w_i, v'_i)^2 \\ &= \binom{\ell-n}{\ell-k} b^{k-n} \prod_{i=1}^{n} (v'_i + v_i'^2), \end{aligned}$$

as required.

(ii) The first equation is proved easily considering Remark 2. To prove the second one, observe that

$$Q_{\ell k} \Lambda^{-1} Q_{\ell k}^{\top} = \Delta_{\ell k} E \Lambda^{-1} E^{\top} \Delta_{\ell k}^{\top},$$

in which the matrix $D = E\Lambda^{-1}E^{\top}$ is a diagonal matrix of the form $D = diag((d_{v'})_{v' \in V'_{\ell,\leq k}})$ with entries $d_{v'} = \frac{1}{\lambda_{|\overline{G}_{v'}|} ||x_{v'}||^2}$ or

$$d_{v'} = \frac{1}{\binom{\ell-|\overline{G}_{v'}|}{\ell-k}^2 b^{\ell-|\overline{G}_{v'}|} \prod_{i \in \overline{G}_{v'}} (v'_i + v_i'^2)}.$$

(iii) Since the columns of $Q_{\ell k}$ are normalized orthogonal eigenvectors, the first identity holds. To prove the second one, using the notation of Remark 2, we begin by claiming

$$A_{\ell k}^{\top} N_{\ell k} = 0.$$

In fact if $y \in \ker(A_{\ell k} A_{\ell k}^{\top})$ then from $A_{\ell k} A_{\ell k}^{\top} y = 0$ we obtain $y^{\top} A_{\ell k} A_{\ell k}^{\top} y = 0$ and $||A_{\ell k}^{\top} y|| = 0$, thus $A_{\ell k}^{\top} y = 0$, which proves the claim. Now, by using the decomposition $P_{\ell k} = [Q_{\ell k}\ R_{\ell k}]$ and the equation $P_{\ell k} P_{\ell k}^{\top} = I$, we obtain $Q_{\ell k} Q_{\ell k}^{\top} +$

$N_{\ell k} N_{\ell k}^\top = I$, hence $A_{\ell k}^\top Q_{\ell k} Q_{\ell k}^\top = A_{\ell k}^\top - A_{\ell k}^\top N_{\ell k} N_{\ell k}^\top$ which yields $A_{\ell k}^\top Q_{\ell k} Q_{\ell k}^\top = A_{\ell k}^\top$ using the above claim. Transposing both sides yields the result.

5.5 The Moore-Penrose Pseudo-Inverse of $A_{\ell k}$

In this section we prove that the matrix $W_{\ell k}$, defined in Equation 5.3 is the Moore-Penrose pseudo-inverse of $A_{\ell k}$. We calculate the entries of this matrix:

$$W_{\ell k} = A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top. \tag{5.24}$$

Proposition 5.5.1 *The matrix $W_{\ell k} = A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top$ is the Moore-Penrose pseudo-inverse of $A_{\ell k}$.*

Proof First note that

$$\begin{aligned}
W_{\ell k} A_{\ell k} W_{\ell k} &= (A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top) A_{\ell k} (A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top) \\
&= A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top (A_{\ell k} A_{\ell k}^\top) Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top \\
&= A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top (Q_{\ell k} \Lambda Q_{\ell k}^\top) Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top \\
&= A_{\ell k}^\top Q_{\ell k} (\Lambda^{-1} Q_{\ell k}^\top Q_{\ell k} \Lambda)(Q_{\ell k}^\top Q_{\ell k}) \Lambda^{-1} Q_{\ell k}^\top \\
&= A_{\ell k}^\top Q_{\ell k} \Lambda^{-1} Q_{\ell k}^\top \\
&= W_{\ell k}.
\end{aligned}$$

Second, we have

$$
\begin{aligned}
A_{\ell k} W_{\ell k} A_{\ell k} &= A_{\ell k} (A_{\ell k}^{\top} Q_{\ell k} \Lambda^{-1} Q_{\ell k}^{\top}) A_{\ell k} \\
&= (A_{\ell k} A_{\ell k}^{\top}) Q_{\ell k} \Lambda^{-1} Q_{\ell k}^{\top} A_{\ell k} \\
&= Q_{\ell k} (\Lambda Q_{\ell k}^{\top} Q_{\ell k} \Lambda^{-1}) Q_{\ell k}^{\top} A_{\ell k} \\
&= Q_{\ell k} Q_{\ell k}^{\top} A_{\ell k} \\
&= A_{\ell k} \quad \text{(by Proposition 5.4.2 (iii))}.
\end{aligned}
$$

Now we can calculate the entries of $W_{\ell k}$.

Proposition 5.5.2 *Let $u \in U_\ell$ and $v \in V_{\ell k}$. Let $P = \{i : 1 \leq i \leq \ell, v_i \neq g, v_i = u_i\}$, $Q = \{i : 1 \leq i \leq \ell, v_i \neq g, v_i \neq u_i\}$, $|P| = p$ (and consequently $|Q| = k - p$). Then the value $W_{\ell k}(u, v)$ is obtained as follows*

$$
W_{\ell k}(u, v) = \frac{\binom{k-\ell}{k-p}}{\binom{\ell}{k}\binom{k}{p} b^{\ell}} \sum_{n=0}^{p} \binom{\ell}{n} (b-1)^n. \tag{5.25}
$$

Equation 5.25 shows that the matrix entries $W_{\ell k}(u, v)$ only depend on the number of mismatches $(k - p)$ between the gapped k-mer, v, and the ℓ-mer, u.

Proof Define the matrix $B_{\ell k}$ by

$$
B_{\ell k} = Q_{\ell k} \Lambda^{-1} Q_{\ell k}^{\top}, \tag{5.26}
$$

so $W_{\ell k} = A_{\ell k}^{\top} B_{\ell k}$. First we calculate the entries of $B_{\ell k}$. For this, note that by (5.20)

and (5.21), for any $y, v \in V_{\ell k}$ we have

$$\begin{aligned} B_{\ell k}(y, v) &= \sum_{v' \in V'_{\ell,\le k}} \nu(y, v')\nu(v, v')d_{v'} \\ &= \sum_{v' \in V'_{\ell,\le k}} \frac{\nu(y, v')\nu(v, v')}{\binom{\ell - |\overline{G}_{v'}|}{\ell - k}^2 b^{\ell - |\overline{G}_{v'}|} \prod_{i \in \overline{G}_{v'}} v'_i(v'_i + 1)}. \end{aligned}$$

Now, fixing $u \in U_\ell$ and $v \in V_{\ell k}$, by using $W_{\ell k} = A_{\ell k}^\top B_{\ell k}$, we have

$$W_{\ell k}(u, v) = \sum_{y \in M_{\ell k}(u)} B_{\ell k}(y, v),$$

which yields

$$
\begin{aligned}
W_{\ell k}(u,v) &= \sum_{y\in M_{\ell k}(u)} \sum_{v'\in V'_{\ell,\le k}} \frac{\nu(y,v')\nu(v,v')}{\binom{\ell-|\overline{G}_{v'}|}{\ell-k}^2 b^{\ell-|\overline{G}_{v'}|} \prod_{i\in\overline{G}_{v'}} v'_i(v'_i+1)} \\
&= \sum_{n=0}^{k} \frac{1}{\binom{\ell-n}{\ell-k}^2 b^{\ell-n}} \sum_{v'\in V'_{\ell n}} \frac{\nu(v,v')}{\prod_{i\in\overline{G}_{v'}} v'_i(v'_i+1)} \sum_{y\in M_{\ell k}(u)} \nu(y,v') \\
&= \sum_{n=0}^{k} \frac{1}{\binom{\ell-n}{\ell-k} b^{\ell-n}} \sum_{v'\in V'_{\ell n}} \frac{\nu(v,v')\nu(u,v')}{\prod_{i\in\overline{G}_{v'}} v'_i(v'_i+1)} \quad \text{(by (5.14))} \\
&= \frac{1}{b^\ell} \sum_{n=0}^{k} \frac{1}{\binom{\ell-n}{\ell-k}} \sum_{t=0}^{n} (-1)^{n-t} \binom{p}{t}\binom{k-p}{n-t} (b-1)^t \quad \text{(by (5.16))} \\
&= \frac{1}{b^\ell \binom{k}{p}} \sum_{n=0}^{k} \frac{1}{\binom{\ell-n}{\ell-k}} \sum_{t=0}^{n} (-1)^{n-t} \binom{k}{p}\binom{p}{t}\binom{k-p}{n-t} (b-1)^t \\
&= \frac{1}{b^\ell \binom{k}{p}} \sum_{n=0}^{k} \frac{\binom{k}{n}}{\binom{\ell-n}{\ell-k}} \sum_{t=0}^{n} (-1)^{n-t} \binom{n}{t}\binom{k-n}{p-t} (b-1)^t \quad \text{(by (5.11))} \\
&= \frac{1}{b^\ell \binom{k}{p}} \sum_{n=0}^{k} \frac{\binom{\ell}{n}}{\binom{\ell}{k}} \sum_{t=0}^{n} (-1)^{n-t} \binom{n}{t}\binom{k-n}{p-t} (b-1)^t \\
&= \frac{1}{b^\ell \binom{k}{p}\binom{\ell}{k}} \sum_{n=0}^{k} \sum_{t=0}^{n} (-1)^{n-t} \binom{\ell}{n}\binom{n}{t}\binom{k-n}{p-t} (b-1)^t \\
&= \frac{\binom{k-\ell}{k-p}}{b^\ell \binom{k}{p}\binom{\ell}{k}} \sum_{n=0}^{p} \binom{\ell}{n} (b-1)^n, \quad \text{(by (5.12))},
\end{aligned}
$$

as required.

Remark 3. Alternatively, one can use an eigendecomposition for $A^\top A$, given by $A_{\ell k}^\top A_{\ell k} = Q_{\ell k}\Lambda Q_{\ell k}^\top$ and obtain $W_{\ell k} = Q_{\ell k}\Lambda^{-1}Q_{\ell k}^\top A_{\ell k}^\top$. Proposition 5.3.2 can be used to show that $z_{v'}^{\ell n}$ defined in Definition 5.3.4 are eigenvectors for the matrix $A^\top A$, where

$x_{v'}^{\ell kn}$ are eigenvectors for $AA^\top$ (corresponding nonzero eigenvalues).

Now we are at the point to prove Theorem 5.2.2:

Proof of Theorem 5.2.2. Considering Proposition 5.5.2 and setting $p = k - m$, the result is obtained.

Corollary 5.5.3 *With notations of the previous Proposition, suppose that $u \in U_\ell$ and $v \in V_{\ell k}$ have exactly m mismatches and define $w'_{\ell km}$ by $w'_{\ell km} = b^\ell \binom{\ell}{k} W_{\ell k}(u, v)$. Then the values $w'_{\ell km}$ satisfy the following recursive equations.*

$$w'_{\ell,k,0} = w'_{\ell-1,k,0} + (b-1)w'_{\ell-1,k-1,0} \tag{5.27}$$

$$w'_{\ell,k,m} = \frac{k-\ell}{k} w'_{\ell,k-1,m-1},\ 0 < m \leq k. \tag{5.28}$$

Proof Since u and v have m mismatches, using notations of Proposition 5.5.2, we have $|Q| = k - p = m$, so $p = k - m$. Hence,

$$w'_{\ell,k,m} = \frac{\binom{k-\ell}{m}}{\binom{k}{m}} \sum_{n=0}^{k-m} \binom{\ell}{n} (b-1)^n. \tag{5.29}$$

Now, if $m = 0$ then $w'_{\ell,k,0} = \sum_{n=0}^{k} \binom{\ell}{n}(b-1)^n$ and (5.27) is easily concluded. To prove the other recurrence relation, suppose that $m > 0$ and observe that by (5.29),

$$\frac{w'_{\ell,k,m}}{w'_{\ell,k-1,m-1}} = \frac{\binom{k-\ell}{m}}{\binom{k-1-\ell}{m-1}} \frac{\binom{k-1}{m-1}}{\binom{k}{m}} = \frac{k-\ell}{k}.$$

This concludes the derivation of the matrix $W_{\ell k}$.

5.6 A Basis for $\mathrm{row}(A_{\ell,k})$

This section gives some additional insight into why the rank of A is less than $\binom{\ell}{k}b^k$, because not all of the equations for the gapped k-mers are independent.

It is easily proved that the matrix $A_{\ell k}$ admits the following block decomposition, where $0\cdots$, for instance, stands for the set of all indexes commencing with 0.

$$A_{\ell,k} = \begin{array}{c} \\ 0\cdots \\ 1\cdots \\ \vdots \\ b-1\cdots \\ \hline g\cdots \end{array} \begin{array}{c} \begin{array}{cccc} 0\cdots & 1\cdots & \cdots & b-1\cdots \end{array} \\ \begin{pmatrix} A_{\ell-1,k-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & A_{\ell-1,k-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & A_{\ell-1,k-1} \\ \hline A_{\ell-1,k-1} & A_{\ell-1,k-1} & \cdots & A_{\ell-1,k-1} \end{pmatrix} \end{array}$$

This decomposition gives some information about the row space of $A_{\ell,k}$. Here we propose an explicit basis for this space, but before this we need to define another matrix.

Definition The matrix $A_{\ell,\leq k}$ is a $(0,1)$ matrix with rows and columns indexed by the elements of U_ℓ and $V_{\ell,\leq k}$, respectively and with entries satisfying $A_{\ell,\leq k}(u,v) = 1$ if and only if u is matchable with v. In other words, $A_{\ell,\leq k}$ is the matrix obtained by combining the matrices $A_{\ell,i}$, $i = 0,\cdots,k$, vertically. The row of $A_{\ell,\leq k}$ indexed by a word $w \in V_{\ell,\leq k}$, is denoted by $\mathbf{r}_w$.

Proposition 5.6.1 *The matrix $A_{\ell,\leq k}$ has the same row space as $A_{\ell,k}$, so* $\mathrm{rank}(A_{\ell,\leq k}) =$

$\sum_{n=0}^{k} \binom{\ell}{n}(b-1)^n$. *Moreover, the set* $(\mathbf{r}_{w'})_{w' \in V'_{\ell,k}}$ *constitutes a basis for* $\mathrm{row}(A_{\ell,k})$.

Proof First we prove that any row of $A_{\ell,n-1}$ is expressible as a summation of some rows of $A_{\ell,n}$. For this, consider a word $w \in V_{\ell,n-1}$ and with no loss of generality, suppose that w begins with g, say $w = gv$ for some $v \in V_{\ell-1,n-1}$. It is easy to observe that

$$\mathbf{r}_{gv} = \sum_{t=0}^{b-1} \mathbf{r}_{tv}. \tag{5.30}$$

Thus, the row space of $A_{\ell,n-1}$ is contained in the row space of $A_{\ell,n}$, and we provide $\mathrm{row}(A_{\ell,\leq k}) = \mathrm{row}(A_{\ell,k})$, so $\mathrm{rank}(A_{\ell,\leq k}) = \mathrm{rank}(A_{\ell,k})$. Based on this, using the fact $\mathrm{rank}(A_{\ell k}) = \mathrm{rank}(A_{\ell k} A_{\ell k}^{\mathsf{T}})$ and Proposition 5.4.1, we obtain

$$\mathrm{rank}(A_{\ell,\leq k}) = \sum_{n=0}^{k} \binom{\ell}{n}(b-1)^n.$$

Now observe that (5.30) can also be written as

$$\mathbf{r}_{0v} = \mathbf{r}_{gv} - \sum_{t=1}^{b-1} \mathbf{r}_{tv} \tag{5.31}$$

and by consecutive applications of this equation, we conclude that for any word $w = 0^m v$ with $m > 0$ and $v \in V'_{\ell-m,k-m}$, the following equation holds

$$\mathbf{r}_{0^m v} = \sum_{x \in \Gamma_b^m} (-1)^{|x|_g} \mathbf{r}_{xv}. \tag{5.32}$$

Note that the words xv in the right, all belong to $V'_{\ell k}$. We conclude that for any $w \in V_{\ell k} \setminus V'_{\ell k}$, $\mathbf{r}_w$ can be written as a linear combination of some rows $\mathbf{r}_{v'}$ with $v' \in V'_{\ell k}$. Now the set of vectors $(\mathbf{r}_{v'})_{v' \in V'_{\ell,k}}$ generates $\mathrm{row}(A_{\ell,\leq k})$ and has the same

number of elements as the dimension of this space, so it is a basis for row($A_{\ell,\leq k}$), as required.

Remark 4. Some properties of our matrix $A_{\ell k}$ are similar to the set inclusion matrix of t-subsets vs. k-subsets; see for instance [102].

5.7 Summary

We have derived a simple form for $W_{\ell k}$, the matrix which maps gapped k-mer counts to ℓ-mer counts. We have shown that this matrix is the Moore-Penrose pseudo-inverse of $A_{\ell k}$ and yields the MMSE estimate for $\mathbf{x}$. This approach can be used to find robust estimates of ℓ-mer frequencies from limited training data, which should reduce overfitting in statistical model learning from genomic sequence data.

One caveat which should be mentioned is that our MMSE estimate of $W_{\ell k}$ has the property that it does not constrain counts of $\mathbf{x}_{MMSE}$ in Equation 2.9 to be positive. In practice we circumvent this difficulty by thresholding the impulse response of the filter that generates the estimated counts from the training sequences. This detail, and application of our method to sequence classification problems, will be discussed in the next chapter.

Chapter 6

Gapped k-mer Support Vector Machine (GSVM) Model

Based on the robust k-mer frequency estimation method described in the previous chapter, in this chapter we will present a novel class of smoothing filters, called gapped k-mer filters, and apply that to estimate k-mer frequencies from limited training data. Then we describe a sequence similarity score (gscore) based on the robust k-mer frequency estimation method. We use the proposed sequence similarity score as a kernel for a support vector machine and build a general sequence classifier. We will use this classifier to classify nucleosome bound and nucleosome free sequences. We also present a novel data structure and algorithm for fast calculation of the kernel matrix. This makes the GSVM method feasible to problems involving large genomewide datasets.

6.1 Introduction

As described in the previous chapter, most analysis of polymeric biomolecules (e.g. protein, RNA, or DNA) at some point requires a model mapping polymer sequence features to functional molecular structures. In the context of DNA sequence properties, sequence similarity via sequence alignment, alignment to known repeat elements, and CpG islands are well known examples. In protein functional and structural studies, amino acid patterns are frequently mapped to structural or functional motifs (such as leucine zippers or phosporylation sites). These descriptions have been fairly successful. However, when modeling DNA-protein interactions, a core process in transcriptional regulation, there is less consensus on what to use as the best description of a DNA sequence (a binding site) bound by a protein (a transcription factor). Many models of this process use a position weight matrix or PWM to describe the DNA binding site [1,88]. Other approaches use oligomers of fixed length, k, commonly known as k-mers, to describe the DNA binding sites. Using k-mers has the distinct advantage that they reflect the discrete space of all possible DNA molecules of length k, while the space of all possible PWMs is continuous and large. We were initially motivated to generalize k-mer methods when our machine learning algorithms based on k-mer frequencies were found to be effective at predicting enhancers and modeling transcription factor binding sites [58]. Other examples of successful bioinformatic applications based on k-mers as sequence features include sequence homology [90,91], protein homology [92], and predictions of cis-regulatory modules [93],

transcription factor binding sites [94, 95], transcription initiation sites [96, 97], and splice sites [98, 99].

When using k-mers, larger k's will resolve larger binding sites and more accurately reflect biological function. For example, some transcription factors (such as ABF1, CTCF, etc.) have relatively long binding sites that cannot be completely represented by short k-mers. So longer k-mers capture more relevant information; however, there is a limitation on the maximum length k which can be effectively used in statistical algorithms. This is an even more significant problem for the very long sequences bound by nucleosomes. Because longer k-mers are more sparsely populated in any finite training sequence set, there is a maximum length k for which the k-mer frequencies can be robustly estimated. Thus in practice, a k is chosen which is a tradeoff between resolving features and robust estimation of their frequencies. To overcome the finite training set size problem, one approach is to employ gapped k-mer frequencies. A gapped k-mer has a length ℓ, and a number of informative columns within that ℓ-mer, k. We refer to the non-informative columns as gaps within a word of length ℓ. For example, the gapped k-mer `ACGG-T-` would match ungapped ℓ-mers `ACGGATG`, `ACCGGTA`, `ACCGATA`, etc. In the case of TF-DNA binding interactions, the gaps within a binding site can represent bases which do not significantly contribute to the sequence specificiy of the TF-DNA interaction. We will show that using gapped k-mers can substantially improve the reliability of the k-mer frequency estimation for a finite genomic training set, because while k-mers become sparsely populated, gapped

k-mers will still have many instances in the training set, and thus their frequencies can be more reliably estimated.

As an example, let us consider the binding site of the well-studied transcription factor, CTCF, whose binding site is shown in Figure 6.1. Table 6.1 shows the list of top scoring gapped 6-mers of length 12 and their number of occurrences in the CTCF bound (positive) and matched control (negative) sets. It can be observed that these gapped 6-mers have a high amount of overlap with each other. And although they are all statistically significant, each gapped 6-mer is present in a subset of the positive and negative sequences and hence a comprehensive model needs to combine the information from several individual gapped k-mers. When combining the gapped k-mers' counts, their overlap should be taken into account. This chapter presents a method for using observed gapped k-mer frequency distribution for all gapped k-mers to estimate the ungapped ℓ-mer frequencies, which are sparsely populated. In chapter 5, we derived an equation for the minimum mean square error (MMSE) estimate for the ℓ-mer frequencies given the frequencies for all gapped k-mers. Here, we present an algorithm to directly calculate this MMSE estimate for the ℓ-mer frequencies from the training set without the need to calculate the frequencies for gapped k-mers. We do this by finding the transformation that maps the ℓ-mer counts in the training set directly to MMSE estimate for the ℓ-mer frequencies. This transformation turns out to be very simple and expressible in terms of $\ell + 1$ coefficients that weigh the contribution of each sequence in the training set to the MMSE estimate based on the

Figure 6.1: **CTCF binding site.** Position Weight Matrix for CTCF is depicted.

number of mismatches. The proposed algorithm is fast and can be easily modified to guarantee positiveness for the estimated ℓ-mer frequencies. We then present an alignment free sequence similarity score based on the ℓ-mer count estimates and give an algorithm to calculate an SVM kernel based on that. Our results show that this model yields robust estimation for ℓ-mer counts for large ℓ and can significantly improve the classification algorithms to classify biological sequences. In particular, we present results for modeling yeast nucleosome positioning data and mammalian enhancers.

Table 6.1: **Top scoring gapped 6-mers for CTCF binding data**

Gapped 6-mer Sequence	Count in positive set	Count in negative set	P-value
CCnnnnGnnGGC	5596	1377	$< 1 \times 10^{-99}$
CCnCnnGnnGnC	5101	1294	$< 1 \times 10^{-99}$
CnnnAGnnGGCn	4784	1132	$< 1 \times 10^{-99}$
CAnnnGnnGGCn	4619	1055	$< 1 \times 10^{-99}$
CCAnnnGnnGnC	4754	1132	$< 1 \times 10^{-99}$
CCAnnnGnnGGn	5169	1373	$< 1 \times 10^{-99}$
CCnnCnGnnGGn	5365	1496	$< 1 \times 10^{-99}$
CCnnnAGnnGnC	4720	1146	$< 1 \times 10^{-99}$
CCnnCnnGnGGn	5184	1399	$< 1 \times 10^{-99}$
CAnnAGnnGnCn	4302	1066	$< 1 \times 10^{-99}$

Table 6.2: **A representative set of 14-mers**

14-mer Sequence	Count in positive set	Count in negative set	Count estimate in positive set	Count estimate in negative set	estimates ratio	PWM -score
CCACTAGGTGGCGC	17	0	3.44	0.10	35.00	17.65
CCACTAGGGGGCGA	2	0	1.91	0.10	19.37	15.66
CCACAAGGGGGCGC	3	0	3.14	0.17	18.16	15.68
ACACTAGGGGGCAC	2	1	1.45	0.14	10.52	14.84
CCATTAGAGGGCGC	2	0	1.52	0.08	18.54	13.00
AGATGGGATCCACC	3	0	0.21	0.20	1.06	-8.29

6.2 Methods

6.2.1 MMSE Frequency Estimation

As described in the introduction, given a limited training set, ℓ-mer counts will become sparse for large word lengths ℓ. Hence, in the previous chapter, we developed a method using gapped k-mers counts to estimate ℓ-mers counts for longer words [103].

Given the length ℓ, and the number of informative columns k, there exist $M = \binom{\ell}{k}4^k$ possible gapped k-mers. Let us denote these gapped k-mers by v_i, $i = 1..M$, and the number of occurrences (counts) for each gapped k-mer by y_i, $i = 1..M$. Then the MMSE count estimate $\hat{x}$ for a given ℓ-mer, u, is given by

$$\hat{x} = \sum_{i=1}^{M} w_i y_i \tag{6.1}$$

where the weight w_i only depends on the number of mismatches, m, between v_i and u, and is given by the following equation:

$$w(m) = \frac{(-1)^m}{b^\ell \binom{\ell}{k-m}} \frac{\ell - k}{\ell - k + m} \sum_{t=0}^{k-m} \binom{\ell}{t} (b-1)^t \tag{6.2}$$

a result which was proved in chapter 5.[1] Here b is the alphabet size and is equal to four for DNA sequences. The above equation would give a non-zero count estimate even for an ℓ-mer that does not have any exact match in the training set.

Direct use of Equation (6.1) for estimating ℓ-mer counts requires calculation of

[1]Equation (6.2) is equivalent to equation (5.10). We have used the binomial identity $\frac{\binom{k-\ell}{m}}{\binom{\ell}{k}\binom{k}{m}} = \frac{(-1)^m}{\binom{\ell}{k-m}} \frac{\ell-k}{\ell-k+m}$ to explicitly show that w_m's are alternatively positive and negative.

the gapped k-mer counts for all M different gapped k-mers, which for large values of ℓ and k, may become computationally impractical. In addition to that, summing up a large set of floating point numbers may result in poor numerical precision. To overcome these issues, we have developed a method to directly compute $\hat{x}$ by using the training set, without calculating the intermediate gapped k-mer frequencies. The proposed algorithm runs in a time linearly proportional to the size of the training set and hence can be applied for larger values of ℓ and k. For this we introduce a class of filters, which we call gapped k-mer filters (or gkm filters for short), described in the following section.

6.2.2 Gapped k-mer Filters

Given a training set, to compute the ℓ-mer count estimates by using Equation (6.1), one should first calculate the gapped k-mer counts, y_i's, in the training set and then use Equation (6.1) to combine the y_i's with a weight corresponding to the number of mismatches, given by Equation (6.2). This is shown schematically in the top row of Figure 6.2. In this figure, a_{ij}'s are the elements of the incident matrix, A, that maps the ℓ-mer counts in the training set to the gapped k-mer counts. The element $a_{ij} = 1$ if gapped k-mer v_i matches with ℓ-mer u_j and is zero otherwise. The w_{ij}'s are the elements of the matrix W mapping gapped k-mer counts to estimated ℓ-mer counts. In chapter 5 we showed that matrix W is the Penrose-Moore pseudo inverse of A. The element w_{ij} only depends on the number of mismatches between

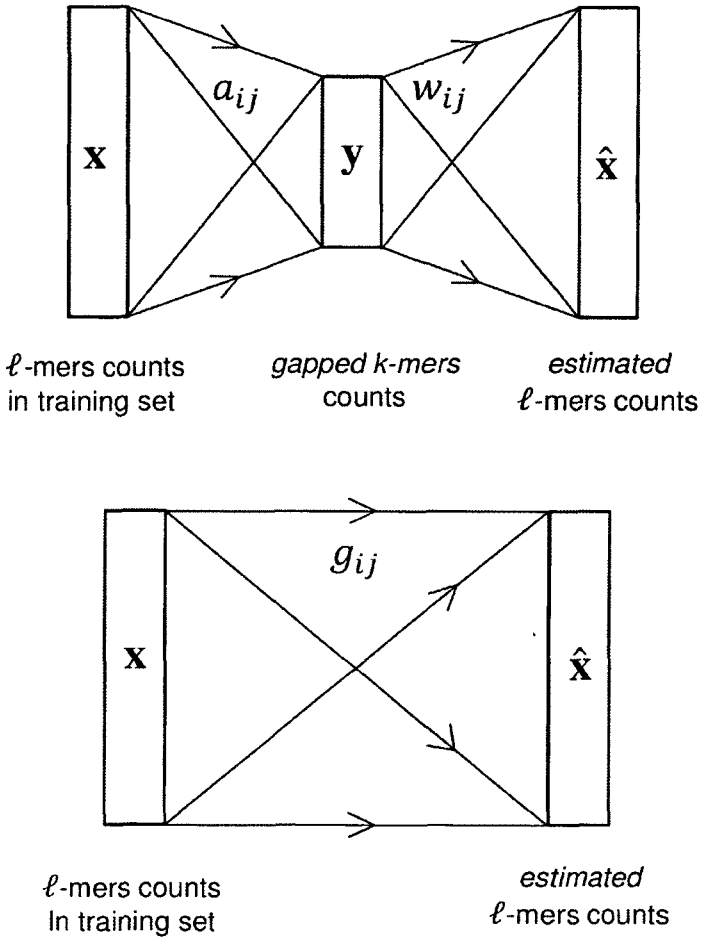


Figure 6.2: **Block diagram for the proposed method.** Top diagram: gapped k-mer counts are obtained from ℓ-mer counts in the training set. Then MMSE ℓ-mer count estimates are obtained from the gapped k-mer counts. The a_{ij}'s are the elements of the incident matrix, A, that maps the ℓ-mer counts in the training set to the gapped k-mer counts. $a_{ij} = 1$ if gapped k-mer v_i matches with ℓ-mer u_j and is zero otherwise. w_{ij}'s are the elements of the matrix W (the pseudo-inverse of A) mapping gapped k-mer frequencies to estimated ℓ-mer frequencies. Bottom diagram: We combine the two mapping matrices A and W and directly calculate the MMSE ℓ-mer count estimates from the ℓ-mer counts in the training set. g_{ij}'s are the elements of matrix G mapping the ℓ-mer counts in the training set to the MMSE ℓ-mer count estimates.

the ℓ-mer u_i and the gapped k-mer v_j and is given by Equation (6.2). Here we show that alternatively, we can combine the two mapping matrices A and W and directly calculate the MMSE ℓ-mer count estimates from the ℓ-mer counts in the training set. This is shown in the second row of Figure 6.2. In this figure, g_{ij}'s are the elements of matrix $G = WA$ mapping the ℓ-mer counts in the training set to the MMSE ℓ-mer count estimates. As we will show below, g_{ij} also only depends on the number of mismatches, m, between the ℓ-mers u_i and u_j. We denote these values by $g_{lk}(m)$ and since the domain and range of this mapping is the same, we call $g_{lk}(m)$ a filter.

To obtain the element $g_{lk}(m)$, that gives the weight for the contribution of an ℓ-mer u_i in the training set to the MMSE count estimate of the ℓ-mer u_j that has exactly m mismatches with u_i, we sum over the contribution of all the gapped k-mers v_τ that match u_i. Note that $a_{ij} = 0$ for the rest of the gapped k-mers. There exist $\binom{\ell-m}{k-t}\binom{m}{t}$ different gapped k-mers that match u_i and have exactly t mismatches with u_j. Figure 6.3 shows how we enumerate all these gapped k-mers. In this figure, the black solid circles denote the m mismatch positions of u_i and u_j, the gray circles denote the $\ell - m$ match positions and the empty dotted circles denote the $\ell - k$ gap positions. For a gapped k-mer to have exactly t mismatches with u_j, there are $\binom{m}{t}$ ways to select the t mismatch positions and $\binom{\ell-m}{k-t}$ ways to select the $k - t$ match positions. Now considering the weight $w(t)$ for the gapped k-mers with t mismatches,

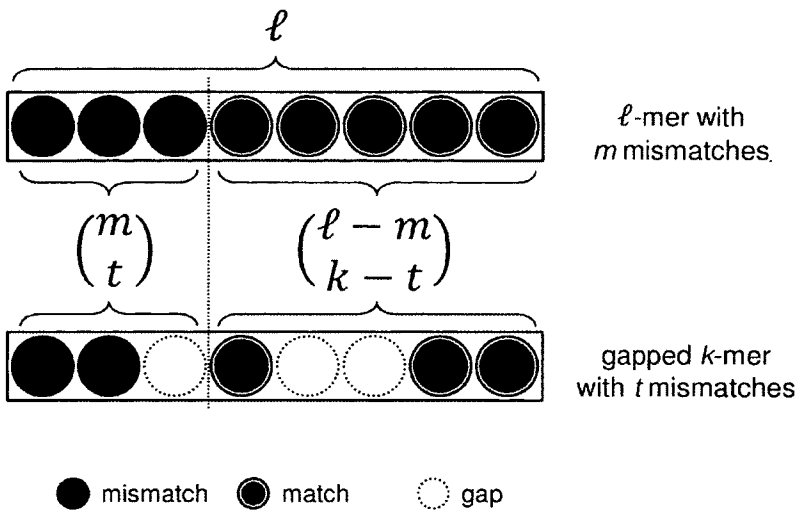


Figure 6.3: **Enumeration of gapped k-mers with exactly t mismatches.** Given the ℓ-mers u_i and u_j, the number of different ways we can construct a gapped k-mer that matches u_i, and has exactly t mismatches with u_j is $\binom{\ell-m}{k-t}\binom{m}{t}$, since there are $\binom{m}{t}$ ways to select the t mismatch positions and $\binom{\ell-m}{k-t}$ ways to select the $k-t$ match positions. In this figure, the black solid circles denote the m mismatch positions of u_i and u_j, the gray circles denote the $\ell - m$ match positions, and the empty dotted circles denote the $\ell - k$ unselected (gap) positions.

the gapped k-mer filter elements, $g_{lk}(m)$ can be obtained as follows:

$$g_{lk}(m) = \sum_{t=0}^{m} \binom{\ell-m}{k-t}\binom{m}{t} w(t). \tag{6.3}$$

There are $\binom{\ell-m}{k-t}\binom{m}{t}$ different ways we can construct a gapped k-mer that matches u_i, and has exactly t mismatches with u_j, by selecting t positions from the m mismatch positions and $k-t$ positions from the $\ell - m$ match positions as explained above and shown in Figure 6.3.

Now using the weights given in Equation (6.3), for any given ℓ-mer, u we obtain the MMSE ℓ-mer count estimate as follows:

$$\hat{x} = \sum_{m=0}^{\ell} N_{tr}(u, m) g_{\ell k}(m) \quad . \tag{6.4}$$

where $N_{tr}(u, m)$ is the number of ℓ-mers in the training set with exactly m mismatches with u.

For large values of ℓ and k, the number of all possible gapped k-mers, $\binom{\ell}{k}4^k$, gets exponentially large and hence this method significantly reduces the CPU time needed to calculate the ℓ-mer count estimates compared to using Equation (6.1). Note that, in contrast to the conventional count, where for any given ℓ-mer, only exact matches in the training set are considered in the count, in the proposed method sequences with mismatches would also contribute to the estimated count with a weight corresponding to the number of mismatches.

Figure 6.4A shows the plots for $g_{\ell k}(m)$ for $\ell = 20$ and various values of k. Each plot is normalized so that weight corresponding to zero mismatch is equal to one. Also Figure 6.4B shows $g_{\ell k}(m)$ for $k = 6$ and various values of ℓ. It can be observed in Figure 6.4A that with a fixed length ℓ, higher values of k result in smaller coefficients for larger mismatches, and therefore less smoothing. For each application, depending on the available training set, one can choose the gkm filter that best fits the application. For the CTCF and P300 examples in this chapter, we used k=min(6,ℓ) and for nucleosome positioning data we used ℓ=10 and k=6. The number of ungapped columns, here 6, represents a trade-off between the amount of training data and the

number of significant columns in a relevant feature in the sequence data.

Also, it can be observed in Figure 6.4 that $g_{\ell k}(m)$ can become slightly negative for large numbers of mismatches, m. This is because in our estimation of the frequencies we did not restrict the frequencies to be positive, and doing so would yield a more complicated expression. (The assumed gaussian distribution allows non-physical negative frequencies to have non-zero probability. A beta-distribution would not have this problem but would introduce offsetting complications.) In cases where the estimated counts are required to be strictly positive, such as when we need to calculate the logarithm or ratios of the estimated frequencies, we truncate the gkm filter $g_{\ell k}(m)$, i.e., for every $m \geq m_0$, we force it to zero, where m_0 is the smallest number of mismatches for which $g_{\ell k}(m_0) < 0$. This will give an approximation of the value of $\hat{x}$ in Equation (6.1), but will guarantee that all the count estimates are non-negative. We have used this approximation for the CTCF, P300 and nucleosome positioning analysis in this chapter.

6.2.3 Naïve Bayes Classifier

We used a similar Naïve Bayes classifier as explained in [104]. In this method, the class prediction is determined by the ratio of likelihood that a sequence is in the positive set relative to the negative set, and features are assumed to contribute independently to this ratio. Hence, for a given DNA sequence of length n, denoted by $\mathbf{s} = \overline{s_0, s_1, \cdots, s_{n-1}}$, class prediction depends on the log-ratio score using $n - \ell + 1$

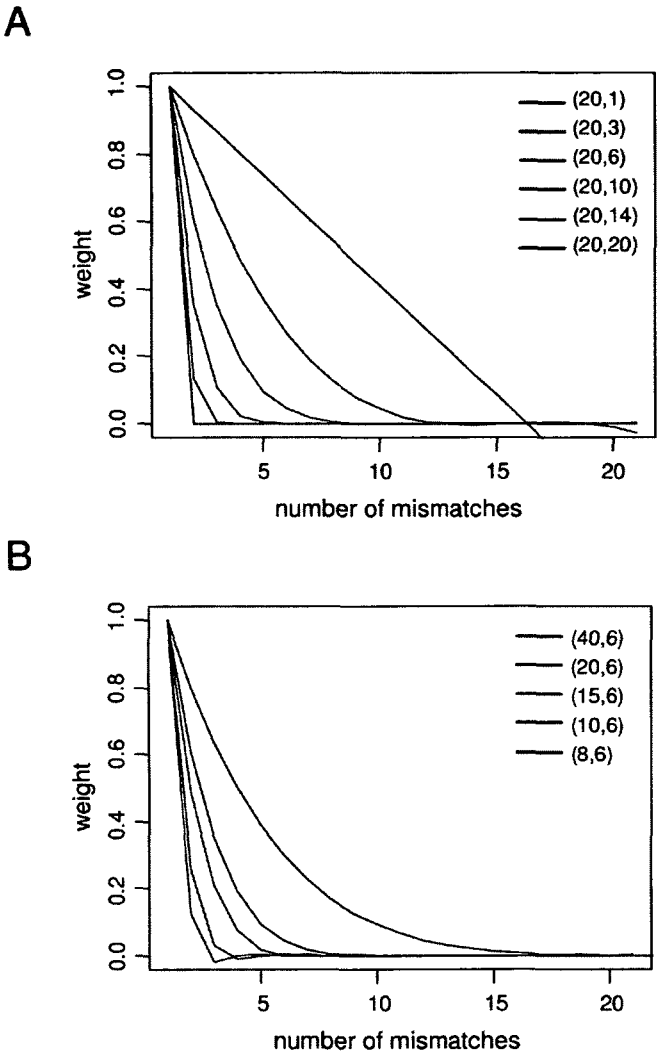


Figure 6.4: $\mathbf{g_{\ell k}(m)}$ **plot.** Plot of the normalized filter function $g_{\ell k}(m)$ for (A) ℓ=20 and various values of k and (B) k=6 and various values of ℓ.

consecutive ℓ-mers' frequencies calculated as

$$\text{logr}(\mathbf{s}) = \sum_{j=0}^{n-\ell} \log \frac{N_p(\overline{s_j, s_{j+1}, \cdots, s_{j+\ell-1}})}{N_n(\overline{s_j, s_{j+1}, \cdots, s_{j+\ell-1}})} \tag{6.5}$$

For Naïve Bayes without gapped k-mer filtering, N_p and N_n are the pseudo-counts for the ℓ-mer, $\mathbf{s}$, in the positive and negative sets respectively. We added 0.5 to the number of occurrences for pseudo-counts. For Naïve Bayes with gapped k-mer filter, N_p and N_n are found using the estimated frequencies given by Equation (6.4) incremented by 0.5 times the smallest positive coefficient of the truncated gkm filter. For any given length ℓ, for each of the sequences in the CTCF test set, we calculated the log-ratio score of Equation (6.5) for every substring of length $n = 15 + \ell - 1$, and assigned the maximum as the sequence score. A window size of 15 was chosen to optimize the performance of the Naïve Bayes without gapped k-mer method for a random subset of the data. For P300 dataset, for each sequence in the test set, we calculated the the log-ratio score of Equation (6.5) for the whole sequence and devided that by the sequence length and used the normalized score for classification.

6.2.4 Support Vector Machine

We have previously developed a support vector machine framework, or "kmer-SVM", for enhancer prediction and have successfully applied to embryonic mouse enhancers [58]. Briefly, our kmer-SVM method finds a decision boundary that maximally discriminates the set of regulatory sequences from random genomic non-regulatory

sequences in the k-mer frequency feature vector space. Here, we extend our original kmer-SVM method to "GSVM" by applying gapped k-mer filtering. Since the straightforward method of generating gapped k-mer frequency vectors becomes quickly impractical due to the exponentially increasing number of k-mers, we propose a new kernel method that directly calculates from sequences the inner product of two gapped k-mer frequency vectors, or gapped k-mer similarity score, as presented in detail in the following section.

6.2.5 Gapped k-mer Similarity Score

Alignment free methods are widely used to compare two sequences of lengths ℓ_1 and ℓ_2 based on the number of occurrences of k-mers. This can be readily extended using ℓ-mer count estimates as following: Given a sequence S, we define the ℓ-mer count estimates vector $f^S = (\hat{x}_1^S, \hat{x}_2^S, \cdots, \hat{x}_N^S)^\top$ where N is the number of all ℓ-mers (4^ℓ in case of DNA sequences), and x_i^S is the count estimate of the ℓ-mer, x_i, in the sequence S using Equation (6.4). Then we define the similarity of two sequences as the inner product of the normalized ℓ-mer count estimate vectors:

$$g - \text{score}(S_1, S_2) = \frac{\langle f^{S_1}, f^{S_2} \rangle}{\|f^{S_1}\| \|f^{S_2}\|} \tag{6.6}$$

where $\|f^S\|$ is the $L2$ norm of the vector f^S and is equal to $\sqrt{\sum_{i=1}^N \hat{x}_i^2}$. The score g-score(S_1,S_2) is always between -1 and 1. Also, for two identical sequences S_1,S_2, we have g-score(S_1,S_2)=1.

For large values of ℓ, direct calculation of g-score using Equation (6.6) is impractical. However, we can use a similar counting approach as we used for obtaining ℓ-mer estimate equation and derive an equation for g-score that does not involve the computation of individual ℓ-mer estimates. We show that the inner product of the two ℓ-mer estimate vectors can be obtained as follows:

$$\langle f^{S_1}, f^{S_2} \rangle = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} c(u_i^{S_1}, u_j^{S_2}) \tag{6.7}$$

where n_1 and n_2 are the number of ℓ-mers in S_1 and S_2, and $u_i^{S_1}$ is the i'th ℓ-mer in S_1 and $u_j^{S_2}$ is the j'th ℓ-mer in S_2 and if u_1 and u_2 have exactly m mismatches then $c(u_1, u_2) = c_m$. Grouping all the ℓ-mer pairs with m mismatches, we can rewrite Equation (6.7) as follows:

$$\langle f^{S_1}, f^{S_2} \rangle = \sum_{m=0}^{\ell} N_m(S_1, S_2) c_{\ell k}(m) \tag{6.8}$$

where $N_m(S_1, S_2)$ is the number of pairs of ℓ-mers $(u_i^{S_1}, u_j^{S_2})$ with m mismatches, and $c_{\ell k}(m)$ is the corresponding weight. We call $N_m(S_1, S_2)$ the mismatch profile of S_1 and S_2. We show that the weight $c_{\ell k}(m)$, denoted in short by c_m, can be obtained as following:

$$c_m = \sum_{m_1} \sum_{m_2} \sum_{t} \binom{\ell - m}{t} \binom{m}{m_1 - t} \binom{m_1 - t}{r} (b-1)^t (b-2)^r g_{m_1} g_{m_2} \tag{6.9}$$

where $r = m_1 + m_2 - 2t - m$, b is the alphabet size and is equal to 4 for DNA sequences. The summations are taken over the range $0, \cdots, \ell$. Figure 6.5 shows how we obtained the equation for c_m. Given two ℓ-mers u_1 and u_2, with m mismatches

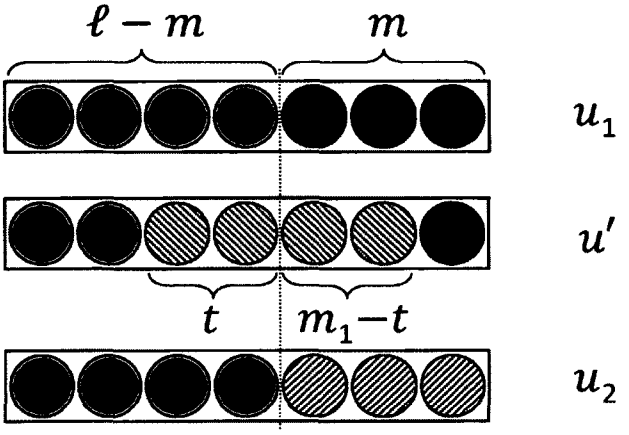


Figure 6.5: **Enumeration of ℓ-mers with $\mathbf{m_1}$ and $\mathbf{m_2}$ mismatches.** Given two ℓ-mers u_1 and u_2, with m mismatches and $\ell - m$ matched places, we want to enumerate the number of all possible ℓ-mers, u', that has m_1 mismatches with u_1 and m_2 mismatches with u_2. For this, we assume that t of the m_1 mismatches are among the $\ell - m$ match positions and $m_1 - t$ of them are among the m mismatch positions. There are $\binom{\ell-m}{t}\binom{m}{m_1-t}$ ways to choose these m_1 positions and $(b-1)^t$ choices for the values of the t mismatches. These t mismatches plus the $m - (m_1 - t)$ unselected mismatch positions are also mismatch for u_2, for the remaining $r = m_2 - (t + m - (m_1 - t))$ mismatches for u_2 there are $\binom{m_1-t}{r}$ ways to select the positions and $(b-2)^r$ ways to select the values. Hence the total number of ℓ-mers, u' with m_1 mismatches with u_1 and m_2 mismatches with u_2, where t of the mismatches of u_1 and u' are among the $\ell - m$ match positions of u_1 and u_2 is given by $\binom{\ell-m}{t}\binom{m}{m_1-t}\binom{m_1-t}{r}(b-1)^t(b-2)^r$.

and $\ell - m$ matched places, we want to enumerate the number of all possible ℓ-mers, u', that has m_1 mismatches with u_1 and m_2 mismatches with u_2. For this, we assume that t of the m_1 mismatches are among the $\ell - m$ match positions and $m_1 - t$ of them are among the m mismatch positions. There are $\binom{\ell-m}{t}\binom{m}{m_1-t}$ ways to choose these m_1 positions and $(b-1)^t$ choices for the values of the t mismatches. These t mismatches plus the $m - (m_1 - t)$ unselected mismatch positions are also mismatch for u_2, for the remaining $r = m_2 - (t + m - (m_1 - t))$ mismatches for u_2 there are $\binom{m_1-t}{r}$ ways to select the positions and $(b-2)^r$ ways to select the values. Hence the total number of ℓ-mers, u' with m_1 mismatches with u_1 and m_2 mismatches with u_2, where t of the mismatches of u_1 and u' are among the $\ell - m$ match positions of u_1 and u_2 is given by $\binom{\ell-m}{t}\binom{m}{m_1-t}\binom{m_1-t}{r}(b-1)^t(b-2)^r$.

Finally, to calculate the norm $\|f^S\|$, we use the following equation:

$$\|f^S\| = \sqrt{\langle f^S, f^S \rangle} \tag{6.10}$$

Using matrix notation, we can show that $g_m = c_m$ if we use the nun-truncated filter $g_{\ell k}(m)$. For that, note that $\langle f^{S_1}, f^{S_2} \rangle = f^{S_1 \top} f^{S_2} = (Gx_1)^\top Gx_2 = x_1 G^\top G x_2$. Given $G = WA$ where W is the Moore-Penrose pseudo-inverse of A, we have $G^\top G = (WA)^\top WA = WAWA = WA = G$. Hence, $\langle f^{S_1}, f^{S_2} \rangle = x_1 G x_2$. This, however, does not hold for the truncated g_m for which we use Equation (6.9) to obtain c_m coefficients.

6.2.6 Cross Validation

Following standard five-fold cross validation procedures, we divided the positive and negative sets into five segments, left one segment out as the test set and used other four segments for training. We repeated for all of the five segments and calculated the mean and standard error of the prediction accuracy on the test set elements.

6.2.7 P-value Calculation

To estimate the statistical significance of a gapped k-mer to be enriched in the positive set we used the hypergeometric distribution to assign a P-value as follows:

$$P = \sum_{i=q}^{\min(r,m)} \frac{\binom{m}{i}\binom{n}{r-i}}{\binom{m+n}{r}} \tag{6.11}$$

where m is the number of ℓ-mers in the positive set, n is the number of ℓ-mers in the negative set, q is the number of ℓ-mers in the positive set that matches with the given gapped k-mer, and r is the total number of ℓ-mers that match the given gapped k-mer. We used the phyper R command [105] and ASA152 C library from StatLib (available online at http://lib.stat.cmu.edu/apstat/) to calculate the above P-value.

6.2.8 CTCF and P300 Datasets

As positive datasets, we used the genome-wide CTCF binding sites in GM12878 cell line [106] and Ep300 binding sites in embryonic mouse forebrain [107], available at

Gene Expression Omnibus (GSE19622 and GSE13845, respectively). For the CTCF dataset, we initially selected the top 5,000 ChIP-seq signal enriched regions, from which we randomly chose 2,500 sites to further reduce the size of the training. For the Ep300 dataset, we defined a new set of the 1,693 400bp sites that maximize the Ep300 ChIP-seq signal within each of the peaks determined by MACS [108] after removing any regions which were more than 70% repeats.

For negative sequences, we found equal numbers of random genomic sequences by matching length, GC and repeat fraction of the corresponding positive set. At each sampling step, we randomly selected a region from the positive set, calculated the length, the GC content and the repeat fraction, and sampled a genomic sequence that matched these properties, and we repeated sampling until we obtained the same number of sequences.

6.2.9 Implementation and Sourcecode

Using truncated gkm filters with m_0 nonzero coefficients for estimating the frequency of an ℓ-mer u, we have to find the number of all ℓ-mers in the training set with at most $m_0 - 1$ mismatches with u. We use a suffix tree to keep the count for all of the ℓ-mers in the training set. Then for a given ℓ-mer u, we traverse all the nodes of the suffix tree within distance m_0 from u and find the count for each number of mismatches. Then we use Equation (6.4) to find the estimated count. For small ℓ's, we store the results in a lookup table to speed up subsequent searches.

For SVM kernel matrix computation, we developed two methods, the exact method and the approximate method. In the exact method, we represent each sequence with a list of ℓ-mers and corresponding count for each ℓ-mer. Then for each pair of sequences, we compute the number of mismatches for all pairs of ℓ-mers and use the corresponding weight c_m to obtain the inner product of Equation (6.8). If the number of unique ℓ-mers in each sequence is of $O(M)$ and the number of sequences is $O(N)$, this algorithm would require $O(M^2N^2)$ comparisons. The naive algorithm for counting the number of mismatches between two ℓ-mers is $O(\ell)$. However, we implemented the mismatch count more efficiently by employing bitwise operations. In summary, using two bits to represent each base $\{A, C, G, T\}$, we used an integer variable to represent t base pairs of the ℓ-mer, therefore using total $\lceil \frac{\ell}{t} \rceil$ integers to represent each ℓ-mer. Then for counting the number of mismatches, we take the bitwise XOR of the integer representations of the ℓ-mers and use a precomputed look-up table to obtain the number of mismatches using the XOR result. This method requires a look-up table of size 2^{2t}. We used $t = 6$ as we found that using a larger t would have negative effect on the performance as the size of the table grows larger. The optimal value of t depends on the processor architecture and amount of cache memory. We also developed a fast approximation method for the SVM kernel which is described in the following section. We have implemented these algorithms and an SVM classifier based on the iterative algorithm described in [109] in C++, and the source code and executable files will become available on our website at http://www.beerlab.org/GSVM.

6.2.10 The Approximate Method (Fast Algorithm for GSVM)

Generally, the coefficient c_m associates larger weights to smaller number of mismatches, m. When using a truncated gkm filter with m_0 nonzero coefficients, $c(m)$ will have $2m_0 - 1$ nonzero coefficients. We have developed a fast algorithm with a parameter m_{max}, that given a set of N sequences, it gives the mismatch profile $N_m(S_i, S_j)$ for every pairs of sequences for up to m_{max} mismatches. We then use Equation (6.8) to obtain the inner products for every pair of sequences. If $m_{max} = 2m_0 - 2$, then this algorithm would give the exact solution. If $m_{max} < 2m_0 - 2$, it will give an approximation of the inner products. For fast calculation of the mismatch profiles, we use a suffix tree structure similar to the mismatch tree explained in [92]. As depicted in Figure 6.6, we use one suffix tree to hold all the ℓ-mers in the collection of all of the sequences. Each node t_i at depth d represents a sequence of length d, denoted by $s(t_i)$, which is determined by the path from the root of the tree to t_i. Each leaf of the tree represents an ℓ-mer, and holds a list of sequences in which that ℓ-mer appeared and the number of times that the ℓ-mer appeared in each sequence. As an example, Figure 6.6 shows the tree that stores all the substrings of length 3 in three sequences S_1=AAACCC, S_2=AAAAA, and S_3=ACC. We initialize the mismatch profile for all the pairs of sequences by zero. Then, starting from the root, we traverse the tree in a depth-first search (DFS) order [87]. When visiting each node t_i, at depth d, we com-

pute the list of all the nodes t_j at depth d for which $s(t_i)$ and $s(t_j)$ have at most m_{max} mismatches. We also compute the number of mismatches between $s(t_i)$ and $s(t_j)$. This list can be computed and updated recursively as we traverse the tree. When reaching a leaf, we increment the corresponding mismatch profile $N_m(S_i, S_j)$ for each pair of sequences S_i in that leaf and S_j in the list. At the end of one DFS traverse of the tree, the mismatch profiles for all of the pairs of sequences are completely determined. To speed up this method, considering the symmetry in the kernel matrix, we only compute the lower triangle of the matrix. Hence, in addition to the maximum mismatch constraint, at each node t_i, we also exclude the nodes t_j in the list that have maxID(t_i)<minID(t_j) where minID(t_i) and maxID(t_j) are the maximum and minimum sequence ID in the subtrees of t_i and t_j respectively and are computed and stored for each node at the time we build the tree.

6.2.11 ROC Curves

To compare the performance of different classification methods, we calculated the area under the receiver operating characteristic (ROC) curve for each classifier. To plot the ROC curves and calculate area under the curves (AUCs) we used the ROCR package [110] in R.

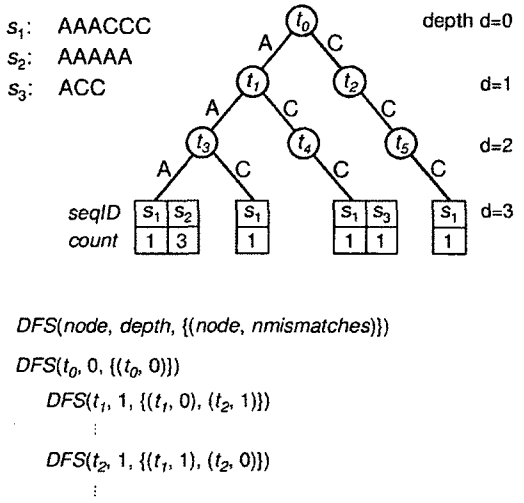


Figure 6.6: **Fast computation of mismatch profiles.** We use a suffix tree in our algorithm. In the above example, ℓ=3 and there are three sequences S_1=AAACCC, S_2=AAAAA, and S_3=ACC. The leafs (nodes at depth $d = \ell = 3$) correspond to 3-mers AAA,AAC,ACC,CCC. The sequence ID and the number of times each 3-mer appeared in each sequence is stored for each leaf. Each node t_i at depth d represents a sequence of length d, denoted by $s(t_i)$, which is determined by the path from the root of the tree to t_i. For example, $s(t_2) = C$ and $s(t_4) = AC$ in this figure. DFS is started at the root node, t_0. When visiting each node t_i, at depth d, we compute the list of all the nodes t_j at depth d for which $s(t_i)$ and $s(t_j)$ have at most m_{max} mismatches. We also compute the number of mismatches between $s(t_i)$ and $s(t_j)$. When reaching a leaf, we increment the corresponding mismatch profile $N_m(S_i, S_j)$ for each pair of sequences S_i in that leaf and S_j in the list.

6.2.12 CTCF Logo

The CTCF logo was generated using the MEME [111] motif discovery software (available online at http://meme.nbcr.net) by using 180 sequences from each of the CTCF bound and CTCF free training sets.

6.3 Results

6.3.1 CTCF and P300 Binding Sequence Modeling

We initially applied our gkm-filtering method to the CTCF binding dataset [106]. As shown in Figure 6.1, CTCF recognizes a specific set of long (up to 20bp) DNA sequences, which severely impair the ability of direct k-mer counting to identify informative sequence features. However, our gkm-filtering method can effectively overcome this limitation and estimate more robust frequencies of those long DNA sequences. To test this hypothesis, we built a Naive Bayes (NB) classifier as described above to discriminate CTCF binding sequences from CTCF free sequences. We also applied the gkm-filtering method to kmer-SVM framework previously developed in [58]. We calculated the area under the ROC curve (AUC) to measure the overall performance of classification, and compare different word lengths (ℓ-mers), with and without the gkm-filtering. As shown in Figure 6.7A, both NB and SVM with the gkm-filtering method perform consistently better than those without the filtering method through-

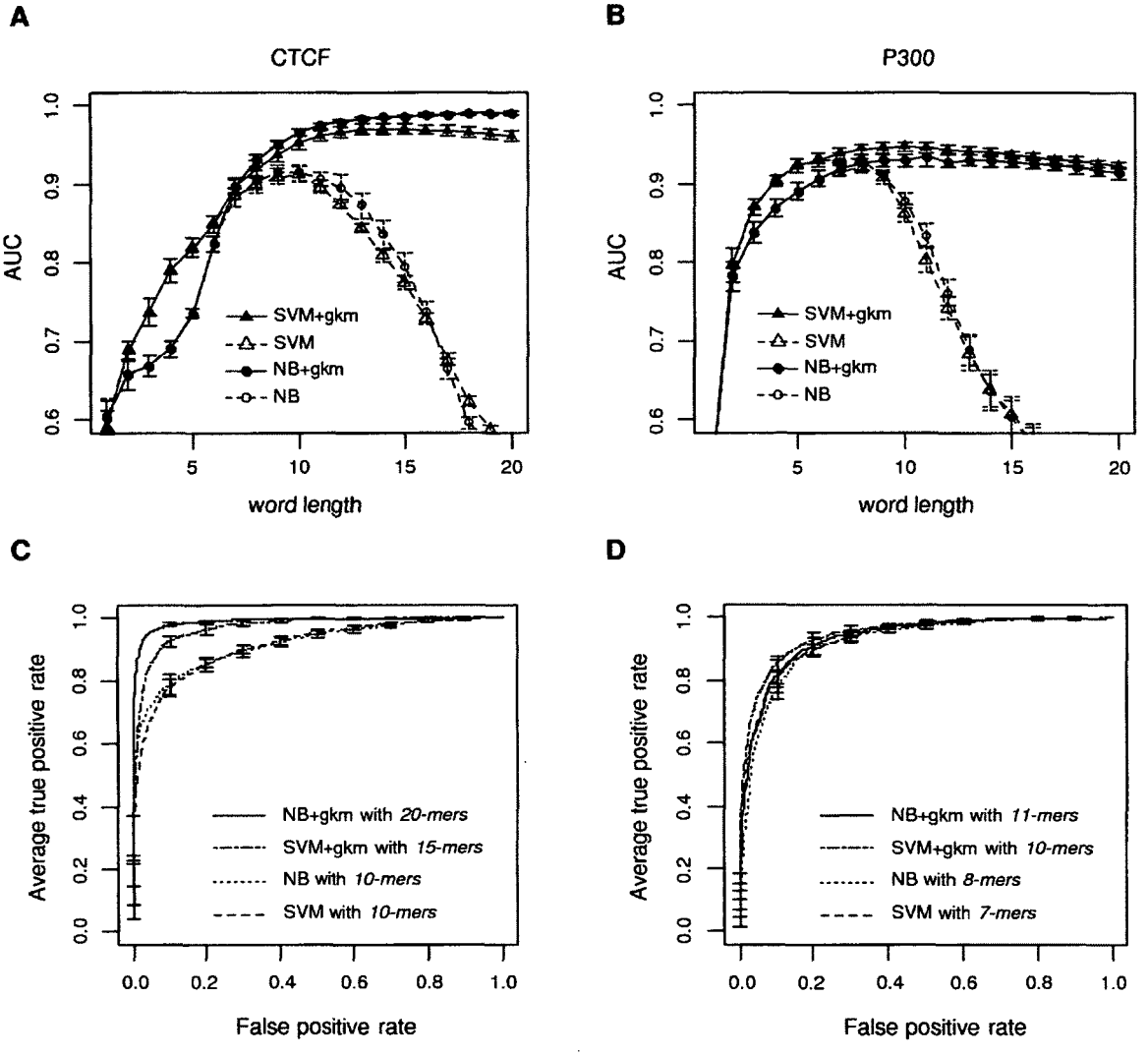


Figure 6.7: **gkm filtering results.** Performance of Naïve Bayes (NB) and SVM classifiers for CTCF and P300 binding sequence modeling with and w/o gapped-kmer filtering. A,B) Area under the ROC curve, C,D) ROC curve for two cases.

out all different word lengths. Most significantly, at longer word lengths ($\ell > 10$), both NB and SVM classifiers using standard k-mer counts greatly suffer from overfitting (or "curse of dimensionality"), but the AUCs of NB and SVM classifiers with the gkm-filtering continue to remain high. The improved performance for CTCF binding site modeling using gkm-filtering is consistent with the notion that the CTCF binding site is long, and that appropriately estimated longer k-mer frequencies as features should improve the classification accuracy. Figure 6.7C shows the best ROC curve of each classification method. NB+gkm achieved the best AUC of 0.992 with ℓ=20, while SVM+gkm achieved significantly lower AUCs of .972 with ℓ=15. A reason for the superior performance of NB compared to SVM is that our NB classifiers assigns the maximum score of all subsequences of length $15+\ell$ to each sequence, while SVM uses the whole sequence when scoring, which includes some redundant neighboring sequences.

CTCF binds to a specific long binding site shown in Figure 6.1. To see whether the gkm filtering can improve models for more degenerate and combinatorial binding sequences, we applied this method to the embryonic mouse EP300 forebrain dataset [107]. We had previously shown that SVM classifiers with the full set of the exact frequencies of k-mers can accurately predict the EP300 binding sites [58]. Figure 6.7B, shows the AUC for NB and SVM classifiers with and without using the gkm filtering. Consistent with the results for CTCF dataset, using gkm-filtering consistently improved the classification performance for P300. The best AUCs of gkm

filtering methods are 0.947 with $\ell = 10$ (SVM) and 0.934 with $\ell = 11$ (NB), whereas those with exact k-mer counting are 0.930 with $\ell = 7$ and 0.921 with $\ell = 8$ (Figure 6.7D). This suggests that longer k-mers contain more information about binding sequences of factors that are interacting with the enhancer and using gkm-filtering makes it possible to better estimate the frequencies for these elements.

Finally, we applied the GSVM method to model the nucleosome positioning data in yeast. The results are given in the next section.

6.3.2 Comparison of GSVM, Simple Context-Based and Phase-Dependent Context-Based Models

To compare the performance of the different methods we developed for modeling nucleosome positioning data, we used the *in vitro* and *in vivo* genomewide nucleosome positioning data from [11] . We sorted genomic positions by nucleosome occupancy signal, and took the top 18000 non-overlapping 147bp sequences (which is about a third of all non-overlapping sequences) as positive set. Also sorted all the positions by negative nucleosome occupancy signal and took the top 18000 non-overlapping 147bp sequences as the negative set. To avoid over-fitting, we separated two of the chromosomes for test (as listed in Table (6.3)) and used the remaining chromosomes in training. For training the models we only used the top 9000 highly bound and

Table 6.3: **Cross validation sets:** To evaluate different models performances, we performed 5-fold cross validation. For each CV set, we separated two chromosomes (about 13.5% of data) as test set and used the remaining fourteen chromosomes for training.

CV set	test chromosomes	total size	% yeast genome
1	chrII, chrXIV	1.60Mb	13.2%
2	chrV, chrXII	1.65Mb	13.7%
3	chrVII, chrVIII	1.65Mb	13.7%
4	chrX, chrXIII	1.67Mb	13.8%
5	chrXI, chrXVI	1.61Mb	13.4%

free sequences, while for test we evaluated the performance over the entire positive and negative sets in the two CVset chromosomes. We used simple context-based model with maximum context size of 12 bp, phase-dependent context-based model, and GSVM model. We also used simple context-based model with maximum context size of 1 for comparison. We used minimum frequency count of 200, and phase bin size of $\frac{\pi}{2}$.

Figure 6.8 shows the results. It can be observed that overall, the sequence based models have a higher accuracy using *in vitro* data compared to *in vivo* data. For *in vivo* data, GSVM and phase-dependent context-based model gave comparable results, with AUC=0.976(0.001) for GSVM and AUC=0.974(0.002) for phase-dependent context-based model. The numbers in parentheses are standard deviation of the five cross validation sets. These two methods outperformed simple context-

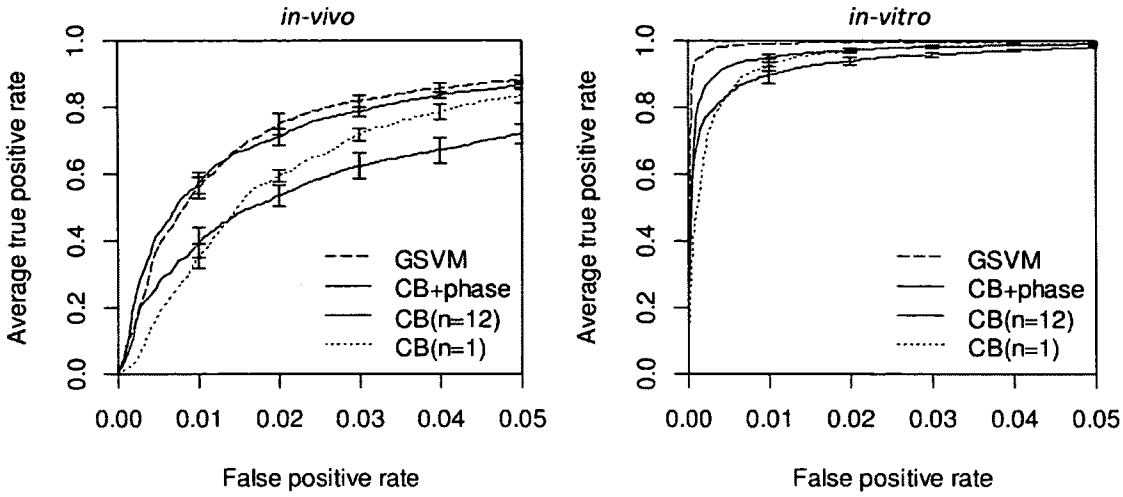


Figure 6.8: Comparison of different nucleosome positioning models.

based method with n=12 (AUC=0.948(0.003) and n=1 (AUC=0.967(0.002)). For *in vitro* data, GSVM gave the best performance (AUC = 0.9996(0.0001)), followed by phase-dependent context-based model (AUC=0.9975(0.0004)). Simple context-based model with n=12, gave AUC=0.9949(0.0009) and context-based model with n=1, gave AUC=0.9964(0.0007).

6.4 Summary and Discussion

6.4.1 Choice of k

In this chapter, we presented an efficient algorithm for computing the ℓ-mer count estimate using gapped k-mer filters. Given a fixed length ℓ, the parameter that

determines how to penalize different number of mismatches is k. Using smaller values of k results in smoother filtering (more tolerance to mismatches), which is more appropriate for more degenerate sequence models or when we have very small amount of training data. As a rule of thumb, the number of ℓ-mers in the training set should be about or greater than b^k. One approach to find an optimal value of k is to try different values of k and use a subset of data as training set and assess the performance on the remaining data points (test set) and take the k that maximizes the performance on the test set. Although the choice of k directly effects the estimated counts, in our simulations we found that the overall performance of the classifier is not very sensitive to changes in k.

6.4.2 Considerations about Arithmetic Error

Direct use of Equation (6.1) for computing count estimates involves a large number of floating point addition operations. Floating point numbers are represented by limited number of digits and arithmetic error may add up to significant amounts when calculating large summations like the one in Equation (6.1). There are some algorithms, such as Kahan compensated summation [112], that significantly reduce this error. However, when using gapped k-mer filters presented in this chapter, we do not need to use such algorithms because, in this method, we first count the number of ℓ-mers with different number of mismatches. This step does not involve any floating point calculations. Then we calculate the weighted sum of the number of mismatches

using Equation (6.4), which only involves ℓ floating point additions. Similarly for the GSVM algorithm, computing the mismatch profiles does not involve any floating point operations, and, in the last step, calculation of the weighted sum only involves ℓ floating point additions.

6.4.3 Generalization of This Work

A simple generalization to the gapped k-mer equation is to use the Gamma function in calculation of the binomial coefficients. In this way, the parameter k does not need to be integer and we can obtain a continuous class of filters defined for any value of k. Also, the weights for different number of mismatches can be optimized for a specific application by using a numerical optimization algorithm such as steepest descent or conjugate gradient. Also when applying the similarity score in an application involving many sequences, instead of using tables to keep the complete mismatch profiles, the algorithm can be optimized using a data structure that only stores the mismatch profiles for sequence pairs with nonzero number of ℓ-mers with at most m_{max} mismatches.

Throughout this chapter, we focused on DNA sequences as features for classification. However in principle, the proposed method can be adopted more generally for any classification or prediction problem involving a large feature set. In general, when the number of features used by a classifier increases, the number of samples in the training set for each point in the feature space becomes smaller, and similar

finite training set issues can occur. One approach is feature selection, which picks a subset of features and builds the classifier only using those features and ignores all the other features. However, usually a unique subset of features that can explain all the variation in the predicted quantity does not exist. An alternative approach would be to use all different k-subsets of the features and then weight them corresponding to the number of mismatches to the test condition in a similar way to gapped k-mers.

Chapter 7

Average Nucleosome Positioning Near Transcription Start Sites Is Dominantly Determined by Transcription Factors

Nucleosome positioning near transcription start sites (TSS) follows a regular pattern on average, consisting of a nucleosome-free region (*NFR*) and well-positioned nucleosomes upstream and downstream of the *NFR*. Both intrinsic DNA-histone interactions and transcription factor-mediated mechanisms are suggested to be responsible for this regular pattern. In the previous chapters, we built sequence based models to predict nucleosome bound and nucleosome-free regions. These models are

mainly based on the intrinsic DNA/histone interactions. Comparing the *in vivo* and *in vitro* average nucleosome occupancy near the TSS shows a regular average pattern that is only present *in vivo*. A recent study shows that similar average pattern can be reproduced *in vitro* by salt gradient dialysis if appropriate amount of ATP and whole cell extract is added [53]. In this chapter, we investigate the regular nucleosome positioning pattern near the TSS. We provide evidence that shows the regular average pattern results from regular positioning of nucleosomes around the TSS in a fraction of yeast genes, and that transcription factor binding sites information is enough to explain this regular pattern. For this, we made a simple model that uses ABF1, REB1 and RAP1 binding sites to reproduce the asymmetric average signal near the TSS with high accuracy. This suggests that the asymmetry in average signal arises from asymmetric distribution of the binding sites with respect to the TSS.

7.1 Introduction

Nucleosome positioning and chromatin remodeling is thought to play a significant role in transcriptional regulation by controlling the accessibility of DNA binding proteins to their DNA binding sites. Recent high-throughput sequencing studies have provided a valuable source of data to evaluate genome-wide nucleosome positioning *in vivo* and *in vitro* in yeast [11, 14]. Although there is a high correlation between the *in vitro* and *in vivo* nucleosome positioning, which shows significant contribution of

sequence dependent DNA- histone interactions, the *in vivo* and *in vitro* nucleosome positioning have some significant differences as well. When the nucleosome occupancy signal is averaged for all genes with respect to transcription start sites (TSS), a clear asymmetric pattern emerges, depicting the underlying regular positioning of nucleosomes around the TSS. Such a regular pattern is not observed *in vitro* [14, 113], suggesting that other factors such as ATP-dependent nucleosome remodeling complexes are needed for the formation of the regular nucleosome patterns near the TSS. [45] showed that ABF1 and REB1 has significant role in forming the nucleosome-free regions (*NFR*) near their binding sites; presumably through interactions with the chromatin remodeling complex *RSC*. [114] showed that many nucleosome-free regions near the TSS are affected by a mutation in *RSC3*, a TF-like subunit of the *RSC* complex. [115] highlighted the role of ABF1 and RAP1. Here to systematically evaluate the contribution of different DNA sequence features, we have looked at the averaged nucleosome occupancy around all 7-mers, and also a comprehensive set of published TF binding sites, and then tried to evaluate the relation between average patterns near individual factors and the averaged nucleosome occupancy near the TSS. Although the average nucleosome patterns around ABF1 and REB1 binding sites are symmetric, an asymmetric distribution of their binding sites near the TSS could explain the overall asymmetric pattern of nucleosome positioning around the TSS. For this, we have developed a simple model that given the binding site information for these transcription factors, and it can reproduce the asymmetric average signal near

the TSS with high accuracy. We have also evaluated some sequence features like poly-T and poly-TA that naturally repel nucleosomes *in vitro*, and shown that although having a significant role in nucleosome positioning *in vivo*, the regular asymmetric average pattern near the TSS cannot be explained merely by these sequence features using the same simple model that we used for transcription factors.

7.2 Average patterns for all 7-mers

We calculated the average nucleosome occupancy patterns for all 7-mers and sorted them by their significance. Figure 7.1 depicts average patterns for three most significant 7-mer patterns: the ATATATA pattern has the deepest nucleosome-free region *in vivo* and *in vitro*. TTTTTTT is known to be enriched in nucleosome-free regions. The next most significant pattern corresponds to the REB1 binding site. Unlike poly-AT and poly-T patterns, REB1 binding sites overlaps with a nucleosome-free region only *in vivo*, suggesting the contribution of the REB1 protein in forming the nucleosome-free region.

We applied hierarchical clustering to cluster significant patterns. Figure 7.1B depicts the average pattern for a selected subset of 7-mers. Some 7-mers, mostly GC-rich (in particular GCnGCnGC), are enriched in nucleosome bound regions, while others, mostly AT-rich, are enriched in nucleosome-free regions. Although each 7-mer generally has a similar pattern *in vivo* and *in vitro*, the pattern is significantly

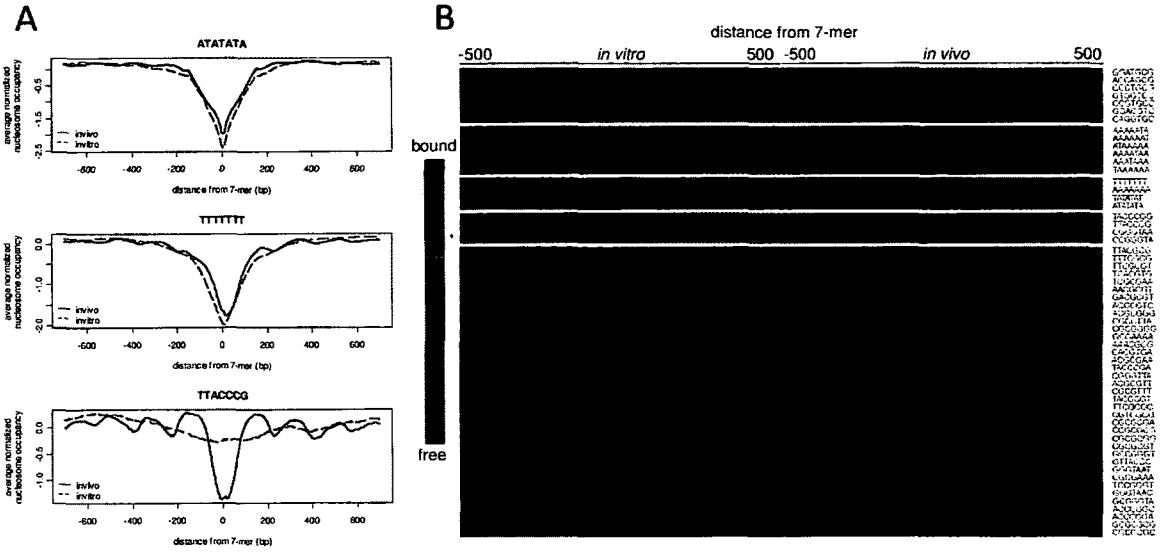


Figure 7.1: **Average Nucleosome Positioning for 7-mers.** (A) The three 7-mers with the most significant *in vivo* patterns. The blue solid curve is the average *in vivo* pattern; the dotted red curve is the *in vitro* pattern. (B) Hierarchical clustering of average patterns near 7-mers shown as a heatmap. On the left is the *in vitro* average nucleosome pattern from -500 to 500 bp from the position of the 7-mer. On the right hand side is the *in vivo* pattern over the same range. 7-mers are sorted by hierarchical clustering so that similar patterns are placed close to each other. A selection of all 16,384 7-mers is depicted.

different for some TF-binding sequences, most significantly TTACCCG (REB1) and TACCCGG(REB1). This difference between the *in vivo* and *in vitro* patterns is also seen —albeit at a lower significance level– for 7-mers containing CGCG or TTACCC elements.

7.3 Average patterns for known motifs

Next, we calculated the average pattern for a list of over a hundred published motifs and used hierarchal clustering to cluster similar patterns. The patterns we obtained were very similar to patterns resulting from 7-mers clustering. The most significant average patterns were corresponding to ABF1 and REB1 (and REB1-related) transcription factors. As can be observed in Figure 7.2, ABF1 has an average pattern very similar to REB1, suggesting that they may be affecting nucleosome positioning through a common mechanism.

7.4 *RRPE* and *PAC* co-occurrence effect

PAC and *RRPE* sites are known to frequently co-occur in yeast genome [116]. The average pattern for all genomic locations that have a *PAC* site shows a nucleosome-free region both *in vitro* and *in vivo*. In order to see whether the *in vitro* nucleosome-free formation is a result of *PAC* sequence, or may have been caused by neighboring *RRPE* sequence, we looked at the average pattern around *PAC* sites that do not

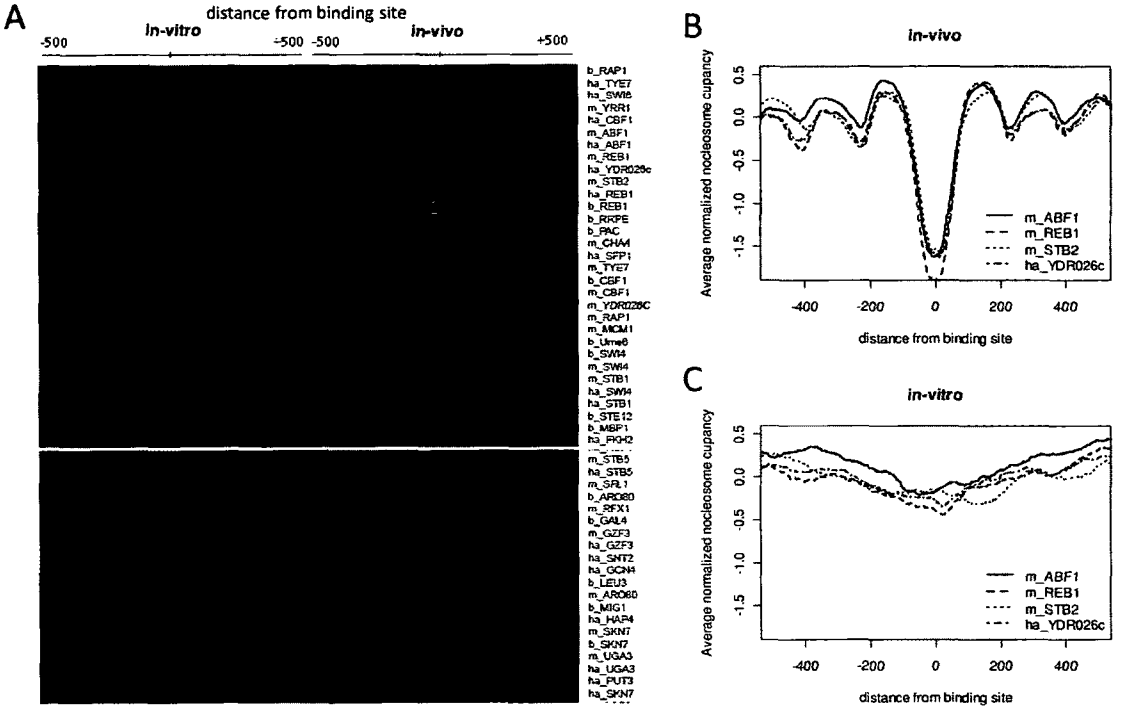


Figure 7.2: **Average nucleosome positioning for selected published motifs.** (A) Hierarchical clustering of transcription factor binding site motifs based on average nucleosome occupancy pattern around instances of the motif. On the left is the *in vitro* average nucleosome pattern from -500 to 500 bp from the binding site. On the right hand side is the *in vivo* pattern at the same range. A selection of the most significant patterns is depicted. (B) *in vivo* pattern for four motifs that had the most significant *in vivo* patterns. (C) *in vitro* average patterns for the same set of TFs.

have a *RRPE* nearby and found a much less significant nucleosome-free region in these cases (Figure 7.3). This suggests that the *PAC* sequence by itself is not enough to form a nucleosome-free region *in vitro*. However, *in vivo* (i.e. in the presence of transcription factors), a nucleosome-free region forms around *PAC* even where there is no *RRPE* site nearby.

7.5 Positional constraint for binding sites relative to TSS for generating nucleosome pattern

The significant differences in the *in vivo* and *in vitro* average pattern for ABF1, REB1, and RAP1 suggest that the binding of these TFs to their binding sites is required for the regular pattern. We looked at the promoters with binding sites for these TFs to see if there is any position constraint to the binding sites. ABF1 and REB1 binding sites are enriched in a window 60 to 170 bp upstream of the TSS. We found that the regular pattern around the TSS preferentially exists only when the TF binding site is within this distance. RAP1 binding sites are more broadly distributed upstream of TSS, but the regular pattern preferentially forms when the binding site is within this distance (Figure 7.4). Moreover, it appears that the position for the first downstream nucleosome is linked more to the position of the TSS than to the

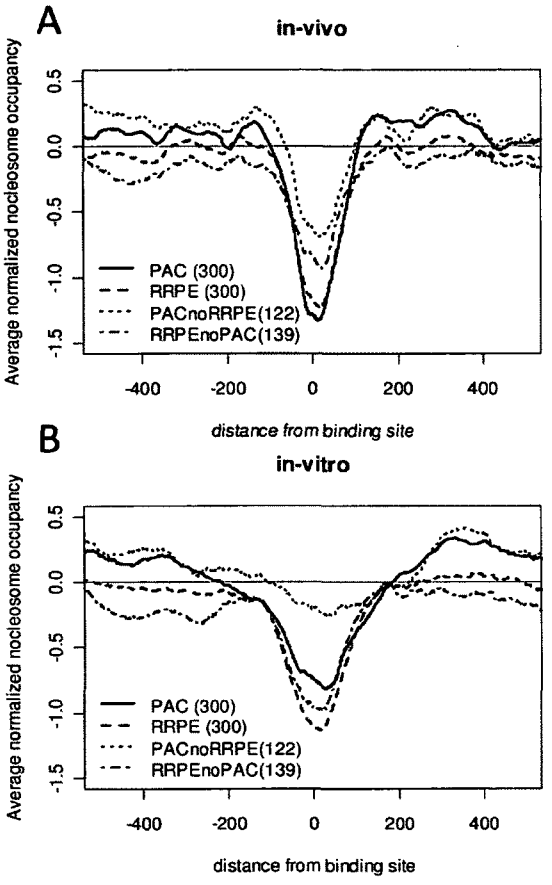


Figure 7.3: ***RRPE* co-localization with *PAC* affects the average *in vitro* pattern for *PAC*.** (A) *in vivo* average nucleosome pattern around the sequence motif for top *PAC* binding sites (blue), top *RRPE* binding sites (red), *PAC* sites that are not near a *RRPE* (magenta), and *RRPE* sites that are not near *PAC* (green). The numbers in parenthesis shows the number of sites. (B) *in vitro* pattern for the same set of sites.

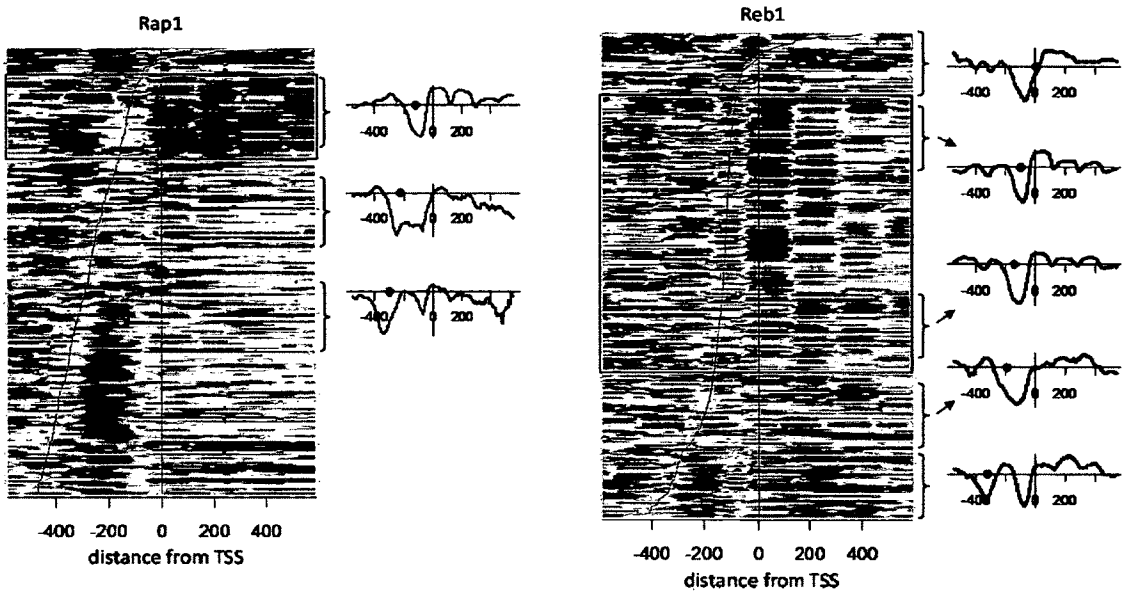


Figure 7.4: **Correlation between motif distance to TSS and the nucleosome pattern.** On the left, the nucleosome pattern for genes containing a RAP1 site near the TSS is shown. Genes are sorted by the position of RAP1 binding site and patterns are aligned by the position of the TSS. The average patterns for three subsets of the genes are plotted next to the heatmap. The blue solid circle depicts the average position of RAP1 binding site for each subset. The blue box on the heatmap highlights the range for which a more regular pattern is observed. On the right, similar analysis is done for REB1 site containing promoters.

TF binding site. These all together suggests that additional factors present at the promoter such as general TFs or RNA polymerase II are required for the regular pattern formation.

7.6 Average pattern around the TSS is not representative of individual genes

Although the average nucleosome positioning data around the TSS for all yeast genes gives a significant pattern, it is an average pattern and does not necessarily represent nucleosome positioning near individual TSSs. We have clustered yeast promoters based on their nucleosome patterns using K-means clustering into four clusters. The average pattern and number of genes for each cluster is shown in Figure 7.4A. It can be observed that a significant subset of genes have no such regular average pattern (cluster 4). We believe that when only looking at the overall average, the nucleosome positioning for these genes may be obscured. Also the genes in the other three clusters are differentially expressed (Figure 7.5B,C). We also found that Cluster 3 is enriched in ribosomal proteins (P-Value $<$ 1E-25 for GO term "cytosolic ribosome"). Moreover, the average expression of genes in each cluster and the distribution of TFs for each cluster are different.

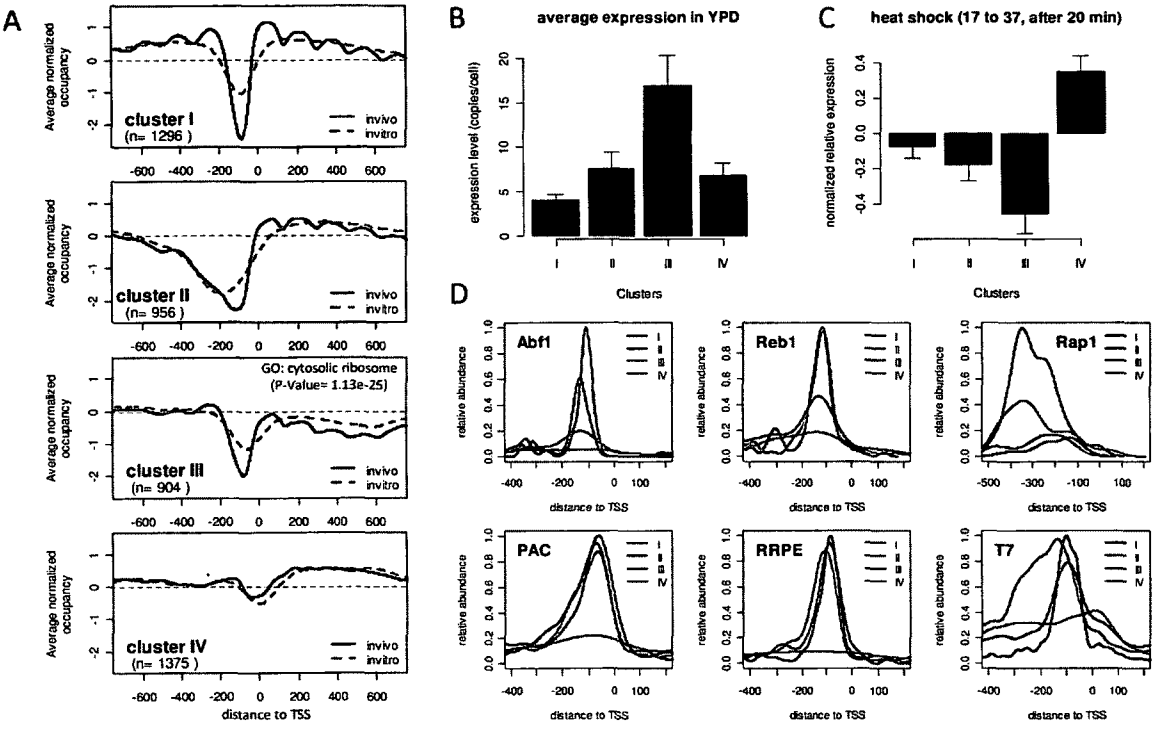


Figure 7.5: **Nucleosome pattern clusters.** (A) Average normalized nucleosome occupancy is plotted for each cluster of genes. The number in the parenthesis reflects the total number of genes in each cluster. Blue solid line is the *in vivo* pattern and red dotted line is the *in vitro* pattern. (B) Average gene expression levels for genes in each cluster. Error bars show the 95% confidence level for mean expression. (C) Relative expression change at heat shock condition averaged for all genes in each cluster. (D) Distribution of different sequence motifs weighted by relative abundance of each motif in each cluster.

7.7 Reconstruction of nucleosome pattern around the TSS using the average pattern for ABF1 and REB1

Significant difference between the average nucleosome pattern *in vivo* and in vitro for some TFs such as ABF1 and REB1 suggests that they may be responsible for the *in vivo/in vitro* differences of the average nucleosome pattern around the TSS. However, the average pattern near the TSS is asymmetric while average patterns around factors such as ABF1 and REB1 are symmetric. In order to see if such TFs are enough to explain the *in vivo/in vitro* difference of average nucleosome patterns around the TSS, we developed a simple model for the nucleosome data by placing the average pattern for ABF1 and REB1 on each ABF1 and REB1 site respectively. We then calculated the average pattern near the TSS for the reconstructed signal. The results are shown in Figure 7.6. It can be observed that the reconstructed pattern using only REB1 and ABF1 binding sites information resembles the average *in vivo* pattern to a great extent. When adding RAP1 sites as well, we obtained an even closer pattern to the average pattern for the real data. We repeated the same method for T7 (TTTTTTT) elements as a control. The reconstructed pattern using T7 did not have the distinguishable downstream nucleosomes and mostly resembled the *in vitro* average pattern. This suggests that information about TF binding sites is sufficient

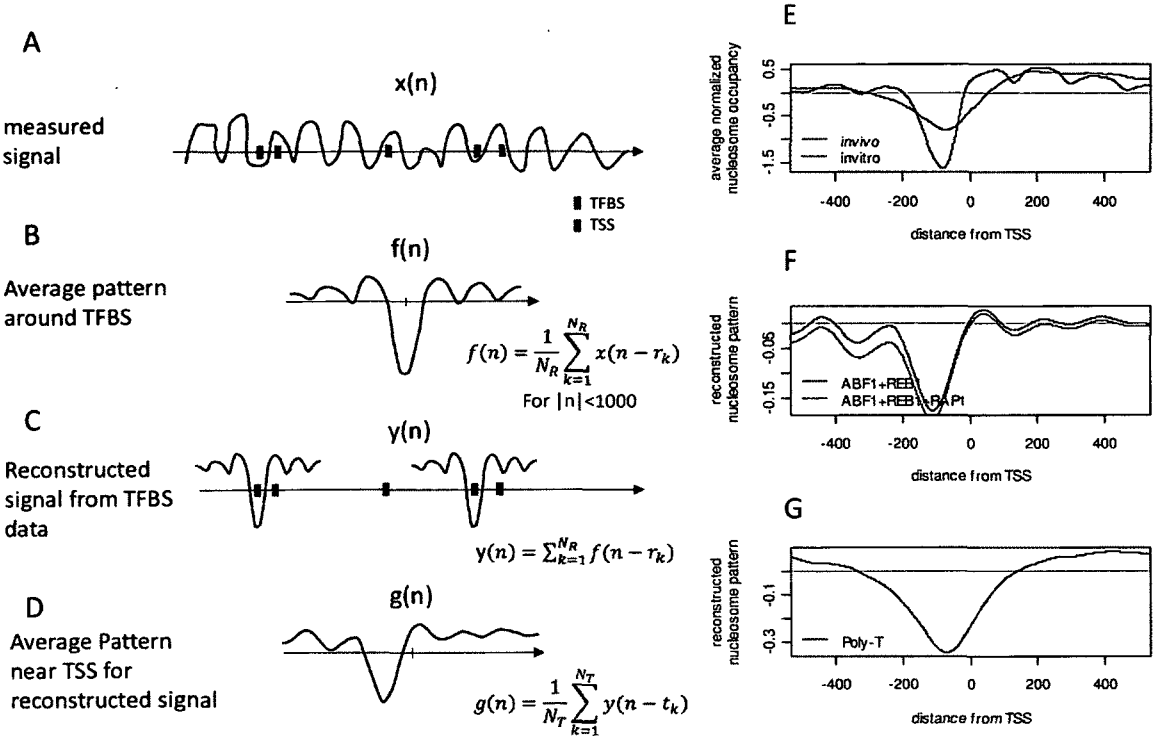


Figure 7.6: **Reconstruction of average pattern near the TSS using TF binding sites information.** (A-D): Schematic of the reconstruction algorithm: average nucleosome pattern around TF binding sites are obtained from the measured nucleosome positioning data. Then the reconstructed signal is obtained by placing the average pattern on the positions of binding sites. Average reconstructed signal around the TSS is then obtained. (E) Average nucleosome pattern near the TSS *in vivo* (blue) and in vitro (red). (F) Reconstructed average pattern using ABF1, REB1 and RAP1 sites. (G) Reconstructed average signal using poly-T.

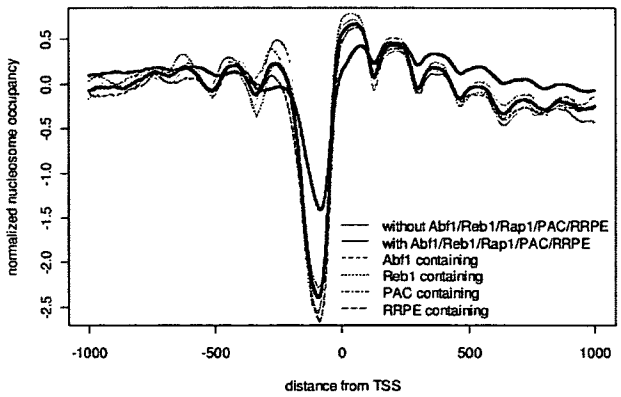


Figure 7.7: **Excluding genes with TF binding sites.** Average nucleosome occupancy is plotted for the set of genes that contain at least one REB1, ABF1, RAP1, *PAC*, or *RRPE* site within 1000bp of TSS (dark blue), compared to the complementary set of genes (magenta). Average patterns for sets of genes containing individual TFs is also plotted using dotted lines.

to explain the regular average pattern around the TSS.

We then checked if removing the genes with sites for TFs would affect the regular average pattern. Figure 7.7 shows the average pattern for genes that contain ABF1/REB1/RAP1/*PAC*/*RRPE* and the complementary set of genes that contain none of these sites. We see that the average pattern, especially for the +1 nucleosome, is significantly reduced by removing genes that have any of these binding sites. The downstream nucleosome occupancy for genes containing these factors is lower than the average for the rest of the genes. One explanation is that histones are not free to move in and fill in the gap formed by shifting the downstream nucleosomes towards the NFR — this is in contrast to the statistical positioning where the only constraint is defined by the nucleosome-free region [14] and is more consistent with

active nucleosome positioning [53].

7.8 Methods

7.8.1 Nucleosome positioning data and genomic sequence

We used nucleosome positioning data from [11] (`http://genie.weizmann.ac.il/pubs/nucleosomes08/nucleosomes08_data.html`). Yeast genomic sequence as of June 2005 was downloaded from the Saccharomyces Genome Database (SGD).

7.8.2 Average pattern for k-mers

For each of the $4^7 = 16384$ different 7-mers, the average pattern was calculated as the average nucleosome occupancy signal around the 7-mer position. The following equation shows the formal definition of the average pattern:

$$P_i(k) = \frac{1}{M_i} \sum_{j=1}^{M_i} f_{chr(s_{ij})}(k - s_{ij}) \tag{7.1}$$

where M_i is the number of incidences for i'th 7-mer, $(1 \leq i \leq 4^7)$, s_{ij} is the genomic position for the j'th incident of i'th 7-mer, $(1 \leq j \leq M_i)$, and $f_{chr(s_{ij})}(k)$ is the nucleosome occupancy corresponding to s_{ij}. Average 7-mer patterns using *in vivo* nucleosome occupancy data were calculated and ranked by their significance using a

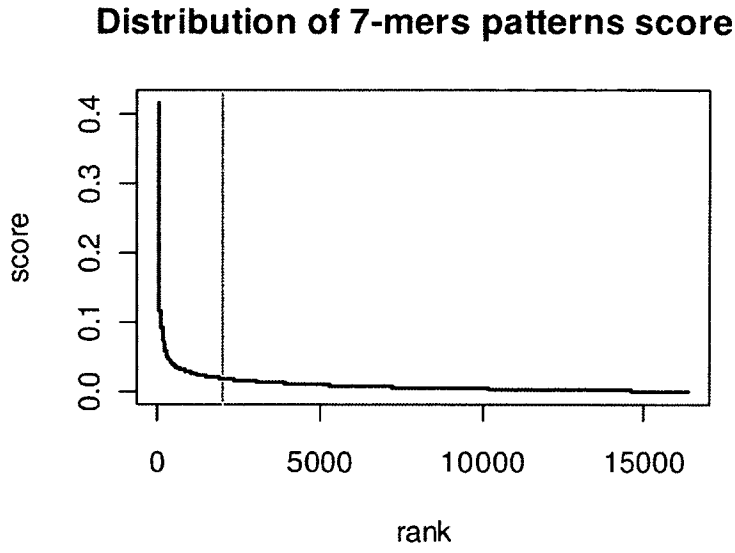


Figure 7.8: Distribution of 7-mers patterns scores

mean square difference measure and the top 2000 7-mers were selected for hierarchal clustering. This was done to make the data size manageable for hierarchal clustering. Figure 7.8 depicts the distribution of scores for all 16384 7-mers. The scores are calculated using following equations:

$$\bar{P}(k) = \frac{1}{16384} \sum_{i=1}^{16384} P_i(k) \tag{7.2}$$

$$S_i = \frac{1}{1001} \sum_{k=-500}^{500} (P_i(k) - \bar{P}(k))^2 \tag{7.3}$$

The top 2000 patterns were applied to hierarchical clustering using hclust command in R. using Euclidean distance and the Ward agglomeration method.

7.8.3 Reconstruction of average pattern

Based on the average nucleosome patterns near Abf1 (and other TFs) binding sites, we reconstructed the signal by placing the average pattern signal at each binding site and then calculated the average signal near the TSS using the reconstructed signal as shown in Figure 7.1. More formally, for a transcription factor with N_R sites, the average pattern is

$$f_(n) = \frac{1}{N_R} \sum_{k=1}^{N_R} x_{chr(r_k)}(n - r_k) \tag{7.4}$$

where $x_{chr(r_k)}(n - r_k)$ is the nucleosome occupancy signal near the kth binding site. Then using the average pattern, we reconstructed the signal as

$$y(n) = \sum_{k=1}^{N_R} f(n - r_k). \tag{7.5}$$

Finally, we used the reconstructed signal and obtained the average pattern near the TSS. The average pattern near the TSS for the reconstructed signal is

$$g(n) = \frac{1}{N_T} \sum_{k=1}^{N_T} y(n - t_k) \tag{7.6}$$

where N_T is the total number of promoters and t_k is the position for the kth TSS.

7.9 Summary and discussion

In this chapter, we developed a simple additive model to reconstruct the nucleosome positioning data based on certain transcription binding sites information. We showed that by only having the information for binding sites of the ABF1, REB1

and RAP1, this model can reproduce average nucleosome pattern around the TSS resembling the *in vivo* average pattern. As a control, we used poly-T binding sites information and also generated the average pattern for nucleosome positioning near promoters that lack a strong Abf1,Reb1,Rap1,PAC, and RRPE site and showed that the regular average pattern significantly reduced for these genes. These observations highlights the role of active mechanisms in specifying *in vivo* nucleosome positioning around the TSS. This also suggests that the asymmetry in the average nuceosome pattern arises from asymmetric distribution of the binding sites with respect to the TSS. We also showed that the regular average pattern is not present in every promoters and there may be multiple mechanisms that specify nucleosome positioinig in different promoters in yeast. We also showed that there is statistically significant relation between the nucleosome pattern near the TSS and gene expression.

Throughout this chapter, we only used data from deep sequencing. For comparison, we also used tiling array nucleosome positioning data [10] and obtained the average pattern near the TSS for each cluster. Results are shown in Figure 7.9. The overall pattern is similar to deep sequencing, but the lower density for downstream nucleosomes is not observed in tiling array data.

An extension to this work would be to build a combined model to integrate the effect of ATP-dependent remodeling factors and intrinsic DNA/histone interactions by quantifying the effect of ATP-dependent factors and the intrinsic binding energies. A simple approach to quantify the relative effect of ATP-dependent factors and

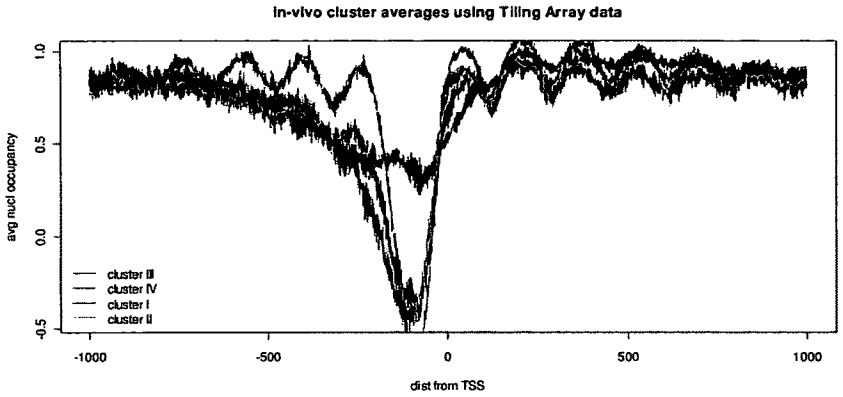


Figure 7.9: Cluster average patterns using tiling array data

intrinsic interactions is to decompose the nucleosome occupancy signal around the TSS to ATP-dependent and intrinsic interactions. Results from the reconstruction algorithm in this chapter suggest that even a simple additive model may be useful to give a first order approximation of the contribution of each mechanism.

Chapter 8

Experimental Validation of Nucleosome Positioning Model

8.1 Introduction

In this chapter we will describe an experimental system we designed to validate the nucleosome positioning models developed in this thesis. Using this system we have made mutations in intergenic regions in yeast and measured the effect of these changes on the expression of the regulated gene. Specifically, we made a set of mutations predicted by our nucleosome positioning models to affect nucleosome affinity to various degrees, and we measured the transcriptional response at the HXT3 locus by inserting a GFP reporter sequence at the 3′ end of the genomic copy of HXT3 gene in haploid yeast. HXT3 is regulated by glucose, and from tiling array data analysis, we

found that there is a nucleosome change in HXT3 promoter upon its activation. We made mutations in that region to make it more nucleosome bound or more nucleosome free according to the sequence based models and developed a method to measure expression level from colonies in a plate using fluorescence microscopy and also used fluorescence plate reader to measure the responsé of HXT3 to different environmental concentrations of glucose for different mutantnts.

8.2 Yeast Hexose Transporter Genes

Glucose is the principal source of carbon and energy in budding yeast *Saccharomyces cerevisiae* [117]. Availability of glucose affects the expression level of many genes in yeast [54, 118, 119]. In yeast, there are 20 genes that are known or likely glucose transporters [120]. Expression of most of these genes are regulated by glucose level in the environment. For example, HXT1 and HXT3 are low affinity (high K_M) glucose transporters which are up-regulated at higher glucose concentrations while HXT2, HXT6 and HXT7 are high affinity (low K_M) glucose transporters which are up-regulated at lower glucose concentrations. By regulating expression of different glucose transporters, yeast optimizes the intake of glucose at different environmental glucose concentrations.

The mechanism by which the HXT genes are regulated has been shown to involve two glucose signaling pathways: Snf3/Rgt2-Rgt1 and Snf1-Mig1 and involves several

transcription factors [121]. None of the current models for transcriptional regulation of HXT genes consider the role of nucleosome positioning. By analyzing the yeast nucleosome positioning data before and after addition of glucose [54] using Group Normalization, we found that there is a significant change in nucleosome positioning near HXT promoters after addition of glucose (see Figure 3.8C) which is consistent with changes in the expression of these genes. We hypothesized that nucleosome remodeling has a role in glucose-dependent regulation of HXT genes in yeast.

8.3 Experiment Design

To test our hypothesis that nucleosome remodeling affects glucose-dependent regulation of HXT genes, we focused on a particular HXT gene, HXT3, for which current models do not fully explain its behavior [120]. Figure 8.1 shows nucleosome occupancy at HXT3 promoter before and an hour after addition of 2% glucose. Current models for HXT3 relate up-regulation of HXT3 to phosphorylation of the repressor protein RGT1 [121–123]. The HXT3 promoter has 10 putative RGT1 binding sites. RGT1 has the consensus binding site (5′-CGGANNA-3′). Multiple glucose signaling pathways converge to phosphorylation of RGT1 upon addition of glucose. Phosphorylated RGT1 does not bind well to its binding site and hence the gene is up-regulated [124, 125]. These models do not consider a role for nucleosome positioning. However, our tiling array data analysis showed a significant change in nucleosome oc-

cupancy at HXT3 promoter after addition of glucose. Moreover, ChIP data from [126] shows that the RSC nucleosome remodeling complex interacts with HXT3 promoter. Two RSC complex subunits RSC3 and RSC30 have binding sites with a core CGCG sequence [114]. The CGCG motif is overrepresented in HXT3 promoter, suggesting that RSC is responsible for nucleosome remodeling at HXT3 promoter. We decided to analyze this region to test our hypothesis that nucleosome positioning has a role in HXT3 glucose-dependent gene regulation and to test whether our sequence based models for nucleosome affinity can predict the effect of mutations in this region on gene expression. We picked this region because we can relatively easily control the glucose in the media and measure the expression of the gene. Moreover, although nucleosome occupancy at HXT3 promoter may be affected by multiple mechanisms, at some glucose concentration threshold the nucleosome is removed. If a mutation changes the energetic cost of DNA bending and nucleosome formation in either direction, we should be able to observe that as a shift in critical glucose concentration at which this region is equally likely to be in the bound or free state.

8.3.1 Random Mutations Selection

Considering the nucleosome profile at HXT3 promoter before and after addition of glucose (Figure 8.1), we targeted two candidate nucleosomes for our further experiments. To choose the right locus to make the mutations, in phase I of our mutagenesis experiments, we deleted 6-8 bp and inserted the T11 sequence ('TTTTTTTTTTT') at po-

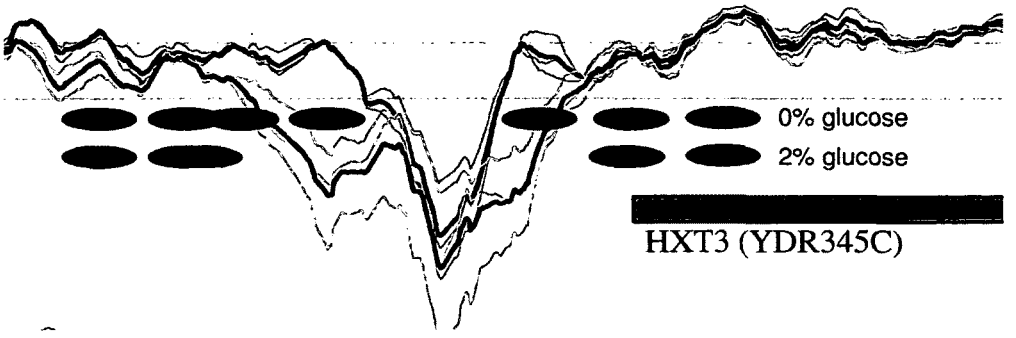


Figure 8.1: Nucleosome organization near yeast hexose transporter gene HXT3 before and 60 minutes after addition of 2% glucose

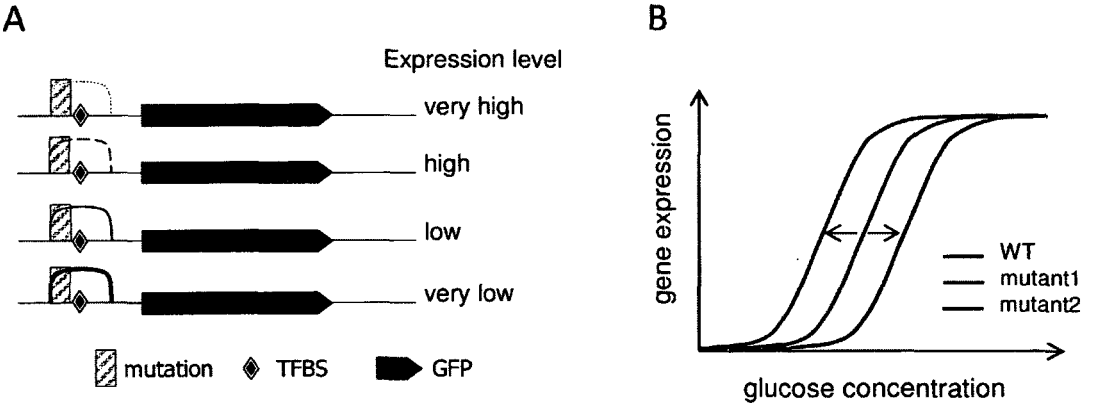


Figure 8.2: Experimental design overview: We made mutations in the HXT3 promoter region according to the sequence-based model to make the region more nucleosome bound or more nucleosome free. We inserted the GFP reporter sequence in 3'end of the genomic copy of the HXT3 gene. A) shows expected expression level for different promoter nucleosome occupancy levels. B) shows expected gene response to glucose for WT and two mutants. We expect that if energetic cost for a nucleosome is increased, the gene should become activated at a lower concentration of glucose (green curve) whereas if the energetic cost for a nucleosome is decreased, it should get activated at higher glucose concentrations (red curve).

sitions 210, 245, 560 and 600 bp upstream of ATG. Poly-T is known to be nucleosome destabilizing. We chose these loci from regions that do not have any overlap with any known sequence elements (e.g. known Rgta binding sites, etc.). We also made mutants with pairs of poly-T sequences inserted at (245,600), (560,600), (245,210), and (560,210) bp upstream of ATG start codon. We found the highest effect for the mutant with poly-T sequences inserted at (245,210). Then we used the phase based context-based model to generate and score 120,000 random mutations at each of the -210 and -245 loci. Then we took all the 10000 combinations of the 100 top scoring random mutations from the -210 locus and the 100 top scoring mutations form the -245 locus, as well as the 10000 combinations of the bottom 100 random mutations of the two loci, plus 100000 random pairs of mutation (randomly drawn from each locus at uniform score spacing) to make a pool of 120000 random mutation candidates. Then we scored these mutations using phase-dependent context-based model and also using simple context-based model with context size (n=4). Figure 8.3 shows the distribution of the scores. It can be observed that there is a high correlation between the scores (Pearson correlation coefficient = 0.98). We manually selected 7 mutations that span a wide range of scores as shown with red circles in Figure 8.3.

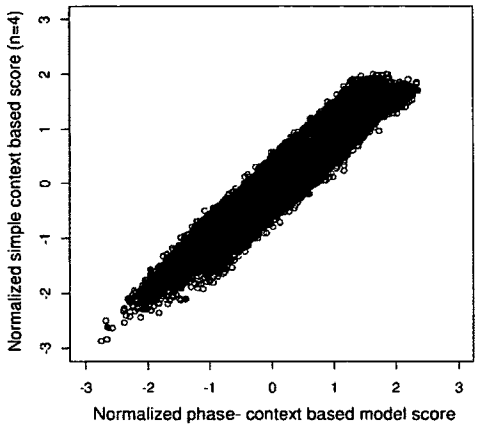


Figure 8.3: **Random Mutations Selection.** Out of a pool of 120,000 randomly generated mutations, we selected 7 mutations that span a wide range of scores. These are shown as red circles in the figure.

8.4 Implementation

To insert GFP and also to implement HXT3 promoter mutations, we first made the construct containing the plasmid with the mutated DNA, homologous flanking regions for recombination, and appropriate selection markers (here URA3 gene for selecting mutants on SD/-URA plate). Then we transformed the DNA to competent yeast cells and selected the mutants on appropriate media (here SD/-URA plates). These steps are explained in detail in the following.

8.4.1 Making Constructs

To measure the HXT3 expression, we inserted the green fluorescence protein (GFP) sequence in 3' end of the HXT3 gene to make a fused protein. For this

we cloned the HXT3 promoter region and inserted that in a plasmid containing yeast URA3 gene (allowing us to select for yeast transformants on -/URA plates), bacterial Ampicillin resistance gene Amp^r (allowing us to select for bacterial transformants on Ampicillin plates), and bacterial origin of replication site *ori* (allowing the plasmid to replicate and get amplified in bacteria). Then we cloned the yeast optimized GFP sequence [127] from a plasmid kindly shared by Brendan Cormack and inserted that at 3′ end of the HXT3 gene before the stop codon. To attach the GFP sequence to HXT3 sequence, we used fusion-based PCR method similar to [128]. We also used similar technique to make two mutants (M725 and M680) as a proof of concept. Using fusion PCR to make mutations requires multiple PCR reactions and gel extraction and is time consuming. For the rest of the mutations (totally 16 mutations) we employed a different technique, using Agilent QuikChange Site-Directed Mutagenesis kit. In this technique Only one pair of primers is needed for each mutation. Below we briefly explain DNA amplification using pYTi plasmid followed by a description of fusion-based PCR and QuikChange Site-Directed Mutagenesis.

8.4.1.1 pYTi Plasmid

To amplify the DNA needed for yeast transformation, we placed it in a plasmid and transformed E. coli, then grew E. coli containing the plasmid in liquid culture, and then extracted the plasmid from E. coli. We made a plasmid, calling it pYTi, based on pHISi, a previous plasmid that we had in the lab and contained URA3 gene,

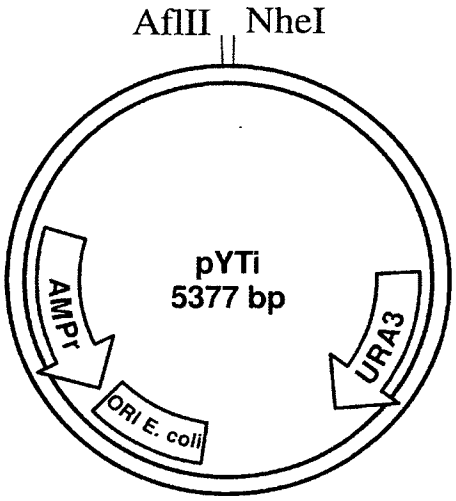


Figure 8.4: **Overview of pYTi plasmid.** We made pYTi by adding AflII and NheI sites and removing HIS gene from pHISi plasmid.

Amp^r gene and bacterial origin of replication site *ori*. We removed the unwanted HIS gene and added new restriction enzyme (RE) sites: AflII and NheI. The choice of these two sites were based on double digestion compatibility, speed, availability in the lab, price per unit of RE, and heat deactivation feasibility in addition to that there is no genomic NheI or AflII sites in HXT locus on chromosome IV. Figure 8.4 shows an overview of the pYTi plasmid plotted using pDRAW32 by AcaClone Software (http://www.acaclone.com).

8.4.1.2 Fusion-based PCR

Figure 8.5 shows the steps required to make a construct with inserted GFP. We employed a method similar to [128]. We started with the plasmid pYTi, yeast genomic

DNA and GFP sequence template. Then we performed two independent PCR reactions using genomic DNA template with primers that have flanking GFP sequences (primer pairs a,b and c,d in Figure 8.5B where b and c have flanking GFP sequences). We also performed a separate PCR reaction to amplify GFP sequence. Then we purified and mixed the PCR products and used it as template for another PCR with a new pair of primers (e and f in Figure 8.5C) to obtain the fused product. These primers also have overhanging AflII and NheI sites and we used them to insert the PCR product in pYTi plasmid by first digesting with restriction enzymes and then ligating with DNA ligase followed by transformation to *E. coli* competent cells. Then we selected the transformed cells on LB plates with ampicilin(AMP). Then made liquid culture and amplified the *E. coli* and subsequently extracted the plasmid, which we then used for yeast transformation. Primers used for these reactions are given in Table (8.1).

Similar technique can be used to implement any mutation, by designing primers with the mutated sequence. We used this technique to implement M725 (a double mutation to convert 'CACCTGGAGGAG' → 'CACCCGGGGGAG' at 725bp upstream of HXT3 gene), and M680 (a double mutation to convert 'CTTTTTCTTGAAAAAAG' → 'CTTTCTCTTGAAGAAAG' at 680bp upstream of the HXT3 gene). Figure 8.6 shows the steps for implementing the mutations using fusion PCR. Primers used for the M680 and M725 mutations are given in Table (8.2)

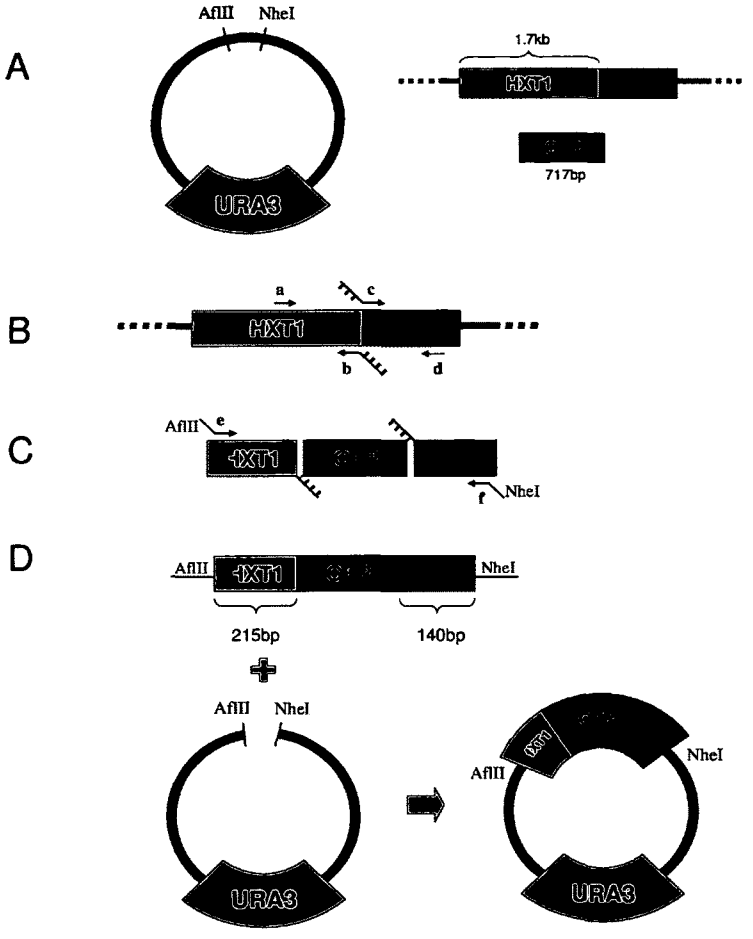


Figure 8.5: **Fusion-based PCR to insert GFP sequence:** A) We started with the plasmid pYTi, yeast genomic DNA and GFP sequence template. B) We performed two independent PCR reactions using genomic DNA template with primers that have overhanging GFP sequences (primer pairs a,b and c,d, where b and c have overhanging GFP sequences). We also performed a separate PCR to amplify GFP sequence. C) We purified and mixed the PCR products and use it as template for another PCR with a new pair of primers (e and f) to obtain the fused product. These primers also have overhanging AflII and NheI sites. D) We inserted the PCR product in pYTi using AflII and NheI sites.

Table 8.1: **List of primers used to insert GFP in 3'end of HXT1, HXT3, and next to ADH1 promoter using fusion based PCR**

Primer	Sequence ($5' \rightarrow 3'$)
GFP(+)	ATGTCTAAAGGTGAAGAATTA
GFP(-)	TTTGTACAATTCATCCATACC
gfp.hxt1.1712(-)	CCAGTGAATAATTCTTCACCTTTAGACATTTTCCTGCTAAACAAACTCTTGTAAAATGG
hxt1.1464+	GCCATCAACTTCTACTACGGTTACGTT
gfp.hxt1.1712(+)	CCATGGTATGGATGAATTGTACAAATAAACTAAACAAGCTCAATATGCATATTTTAATG
hxt1.1883(-)	GAAATTTCTTGTTCCTTGGTGAAGGTC
AflII.hxt1.1496+	TATATACTTAAGGGCTGTATGGTTTTCGCTTACTTT
NheI.hxt1.1851(-)	TATATAGCTAGCAAAGAAAAGAACATCTGTTTATG
HXT3-3':GFP(+)	ATGGATGAATTGTACAAATAATTTCACGCTAAACCGTAG
HXT3-3':GFP(-)	ATGGATGAATTGTACAAATAATTTCACGCTAAACCGTAG
HXT3-3'(+)	TTCTGTTGGTGTCACCAGACT
HXT3-3'(-)	ACGTTCCTAGCAACAAGAGGA
AflII.HXT3-3'(+)	TATATACTTAAGATGGTGAAGGTAATGGTTCAT
NheI.HXT3-3'(-)	TATATAGCTAGCACGATTGTCGGGAGTAAAGAT
ADH1PromOutr(+)	GAATAATTCTTCACCTTTAGACATTTTAGAAGTGTCAACAACGTATCT
ADH1:GFP(+)C	GAATAATTCTTCACCTTTAGACATTTTAGAAGTGTCAACAACGTATCT
ADH1DSOutr(-)	ATGCACGTATACACTTGAGTA
GFP:Adh1DS(+)	CATGGTATGGATGAATTGTACAAATAAGCGAATTTCTTATGATTTATG
AflII.ADH1Pr(+)	GAATATCTTAAGGGTTGACTACATCACGATGAG
NheI-159k(-)	ACTAATGCTAGCTATGTATTCATATCTCAAGAT

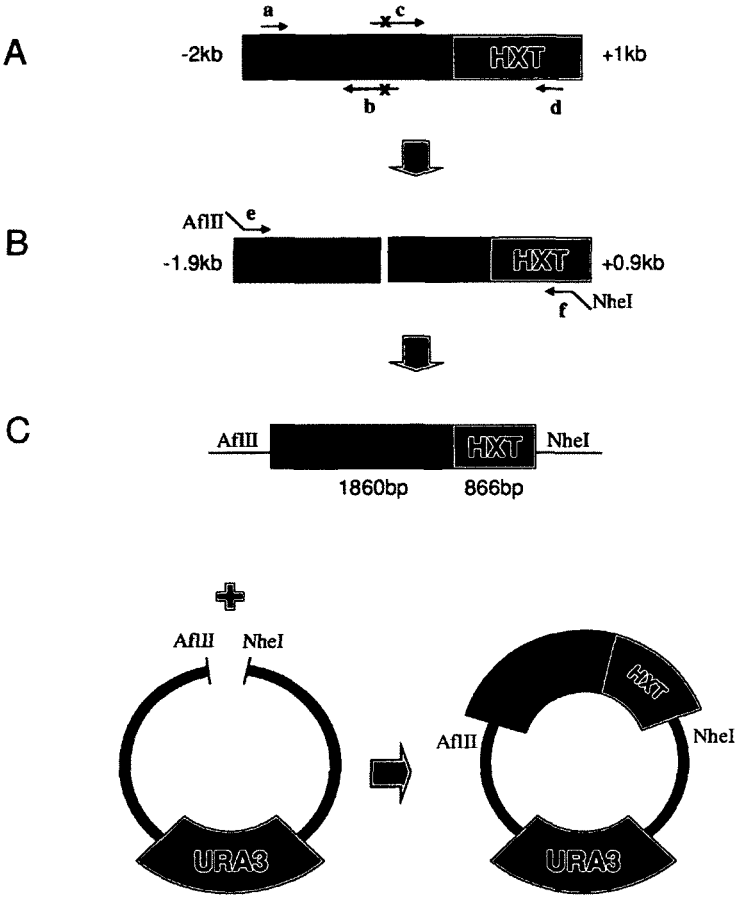


Figure 8.6: **Fusion-based PCR to implement mutations in HXT3 promoter region:**A) Two independent PCR reactions are performed using primers with overlapping primers containing the mutation region. B) PCR product from previous PCR reactions are purified and mixed and used as template for another PCR reaction with primers that have overhanging restriction enzyme site for plasmid insertion. C) AflII and NheI sites are used to insert the fusion PCR product into pYTi plasmid.

Table 8.2: **List of primers used in HXT3 proof of concept mutations.**

Primer	Sequence ($5' \rightarrow 3'$)
Hxt3-736M(+)	GAGATAACACCCGGGGGAGG
Hxt3-714M(-)	GCTCCTCCCCCGGGTGTTAT
Hxt3-688M(+)	CTTTCTTTCTCTTGAAGAAAGAAAA
Hxt3-664M(-)	TTTTCTTTCTTCAAGAGAAAGAAAG
MRgta(+)	CATCGTTCCTGTCACGTAGGATCCCCGATGTTGGTATTTCCGGAGGTCA
MRgta(-)	GAAATACCAACATCGGGGATCCTACGTGACAGGAACGATGACCGGG
Pal40(+)	GGAGCAATGAAATGAAAGCTTGTAAGATGAGCTATTCGCGGAACATTC
Pal40(-)	GCGAATAGCTCATCTTACAAGCTTTCATTTCATTGCTCCTCCTCCAGGT

8.4.1.3 QuikChange Site-Directed Mutagenesis

The fusion based PCR method described above requires at least three PCR reactions and gel extraction for purification of the PCR products. For the rest of the mutations, we used QuikChange Lightning Site-Directed Mutagenesis kit (Agilent) [129]. In this method, one set of divergent primers is designed and the whole plasmid is amplified using a single PCR reaction. Then the PCR product is digested by DpnI restriction enzyme which will digest the methylated template DNA. Then the linear PCR product is transformed to *E. coli* which has enzymes to recircularize the DNA. Finally the plasmid is extracted from *E. coli* and verified by sequencing. We followed the Agilent QuikChange Lightning Site-Directed manual, except for that we used half the amount of the enzymes and performed $25\mu L$ PCR reactions instead

of the $50\mu L$ reactions suggested by the protocol. Moreover, in some cases, we used competent cells made in house instead of the Gold Ultracompetent cells provided by the kit. In those cases we ran PCR for 18 cycles (instead of 16 cycles for gold cells) to compensate for the lower efficiency of our cells. All the mutated plasmids and DNA extracted from mutant yeast strains were verified and validated by DNA sequencing. Finally, the method originally provided by Agilent does not support mutation of more than two bases so we developed our own method to design primers for mutagenesis to implement substitutions of longer lengths. We present a description of the primer design algorithm below.

8.4.1.3.1 Primer Design for mutagenesis

In this section we describe how we designed primers to delete/insert some nucleotides. The idea is to choose a long enough homologous region so that both the PCR step and the recircularization work efficiently, but the homologous region should not be overextended to avoid nonspecific binding. Considering these factors, we came up with the following design steps (Figure 8.7) that we followed to implement the 16 mutations most of them with deletion of 8 bp and insertion of 14 bp.

We used OligoCalc (http://www.basic.northwestern.edu/biotools/oligocalc.html) [130] to predict melting temperature for primers. The design steps are as follows:

- Step1: *ab* and *de* calculation: starting the nucleotide flanking the inserted sequence on each strand, find the minimum number of nucleotides that satisfies

the following criteria:

- Salt adjusted melting temp > 60° C
- Nearest Neighbor melting temp > 53° C
- Last base (3′ end) is a C or G

- Step2: *bcd* calculation: Take the sequence *abcde*, and remove as many nucleotides as possible from both sides (preferentially in a balanced way) to obtain *bcd*, such that the following conditions remain satisfied for *bcd*:
  - Salt adjusted melting temp > 74° C
  - Nearest Neighbor melting temp > 66° C
- The forward primer is *bcde* and the reverse primer is the reverse complement of *abcd*.

All the 16 designed mutations worked, but we also got 2 nonspecific products which were screened out by sequencing.

The list of primers designed and used to implement the mutations is given in Table (8.3).

8.4.2 Yeast Transformation

After preparing the construct containing GFP and the ADH1 promoter (similar procedure for HXT3 and HXT1), we linearized the vector by restriction enzyme di-

Table 8.3: **List of primers used in HXT3 mutagenesis experiment:**

Primer	Sequence ($5' \rightarrow 3'$)
M600pt+	CGTTTGCATCTTCTTGCAAGCTTTTTTTTTTTTTCAATAGTTCGGTAATATTAACGG
M600pt-	CCGAACTATTGAAAAAAAAAAAAAAGCTTGCAAGAAGATGCAAACGAGCTAG
M560pt+	GTAATATTAACGGATACCTTTTTTTTTTTTTCTAGATCCCCTAGTAGGCTCTTTTCAC
M560pt-	GAGCCTACTAGGGGATCTAGAAAAAAAAAAAAAAGGTATCCGTTAATATTACCGAACT
M245pa+	GCCATTATAATGACTGTACAAAAAAAAAAAACTAGTGGAGAAAGAAACAACTCAATAACG
M245pa-	GAGTTGTTTCTTTCTCCACTAGTTTTTTTTTTTGTACAGTCATTATAATGGCGGTC
M210pa+	GAAACAACTCAATAACGATAAAAAAAAAAAGCTTGGGGGCCCACTCAAAAAATC
M210pa-	GTGGGCCCCCAAGCTTTTTTTTTTTATCGTTATTGAGTTGTTTCTTTCTCC
m245A+	CCATTATAATGACTGTACAAATATCTTCTAAGCTTGGAGAAAGAAACAACTCAATAACG
m245A-	GTTGTTTCTTTCTCCAAGCTTAGAAGATATTTGTACAGTCATTATAATGGCGGTC
m245B+	GCCATTATAATGACTGTACAAAAAAAATGTACTAGTGGAGAAAGAAACAACTCAATAACG
m245B-	GAGTTGTTTCTTTCTCCACTAGTACATTTTTTTTGTACAGTCATTATAATGGCGGTC
m245C+	GCCATTATAATGACTGTACAAAAGCTTTTTTTTTTTGGAGAAAGAAACAACTCAATAACG
m245C-	GAGTTGTTTCTTTCTCCAAAAAAAAAAAGCTTTTGTACAGTCATTATAATGGCGGTC
m245F+	GACTGTACAATGGACCACGCTAGCTGGAGAAAGAAACAACTCAATAACG
m245F-	TCTTTCTCCAGCTAGCGTGGTCCATTGTACAGTCATTATAATGGCGGTC
m245G+	TGACTGTACAAGACCATCTAGATGTTGGAGAAAGAAACAACTCAATAACG
m245G-	GAGTTGTTTCTTTCTCCAACATCTAGATGGTCTTGTACAGTCATTATAATGGCGGTC
m210a-	GGCCCCCAACATTGCTGAGGGATATCGTTATTGAGTTGTTTCTTTCTCC
m210a+	ACTCAATAACGATATCCCTCAGCAATGTTGGGGGCCCACTCAAAAAATC
m210b-	GTGGGCCCCCAAAAAAAAAAAAGCTTATCGTTATTGAGTTGTTTCTTTCTCC
m210b+	GAAACAACTCAATAACGATAAGCTTTTTTTTTTTTGGGGGCCCACTCAAAAAATC
m210c-	GTGGGCCCCCAACAAAAAAGTGGGATATCGTTATTGAGTTGTTTCTTTCTCC
m210c+	CAACTCAATAACGATATCCCACTTTTTTGTTGGGGGCCCACTCAAAAAATC
m210d-	GTGGGCCCCCAAGCTTTTTTTTTGTGATCGTTATTGAGTTGTTTCTTTCTCC
m210d+	ACTCAATAACGATCACAAAAAAAAAGCTTGGGGGCCCACTCAAAAAATC
m210e-	GAGTGGGCCCCCAAAAAATTTTTTTGATATCGTTATTGAGTTGTTTCTTTCTCC
m210e+	AGAAACAACTCAATAACGATATCAAAAAAATTTTTTGGGGGCCCACTCAAAAAATC
m210f-	GCCCCCAAGGTGGTGGATATCCATCGTTATTGAGTTGTTTCTTTCTCC
m210f+	CAACTCAATAACGATGGATATCCACCACCTTGGGGGCCCACTCAAAAAATC
m210g-	GTGGGCCCCCAATGGATCCAGATCCAATCGTTATTGAGTTGTTTCTTTCTCC
m210g+	CAATAACGATTGGATCTGGATCCATTGGGGGCCCACTCAAAAAATC

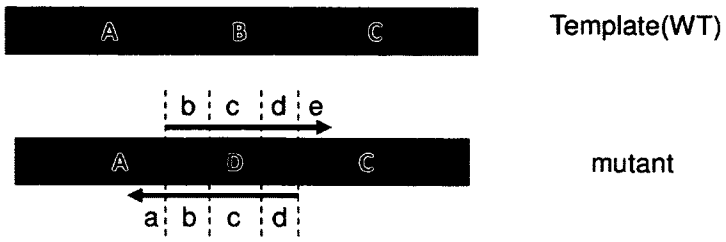


Figure 8.7: **Primer Design for mutagenesis:** Schematic depiction of WT (top) and mutant (bottom) sequences with deletion (B), insertion (D), and unchanged sequences (A,C). B and D may vary in size from 0 to 20 bp. We start by determining the homologous regions for PCR, by choosing the length for the downstream sequence of the inserted sequence '*de*' and '*ab*'. Then we choose the overlapping region for recircularization and select the upstream sequence based on that. The final primers would be '*bcde*' and RevCompl('*abcd*').

gestion and transformed the linear DNA to yeast competent cells that are lacking URA3. The starting cell line that we used was BY4705 (MATα met15Δ0 ura3Δ0 trp1Δ63 leu2Δ0 his3Δ200 lys2Δ0 ade2Δ0) kindly shared by Ed Davis (Boeke lab). We made competent cells using a lithium acetate (LiAc)/polyethylene glycol (PEG) based method [131] and transformed linearized plasmid containing the URA3 gene. We then selected the mutants on SD/-URA plates. Then we checked the colonies under fluorescence microscope and selected a green colony and transfered that to FOA plate (SD plates containing 5-Fluoroorotic Acid). FOA selects against URA3 [132], so the URA3 will hop out as shown in Figure 8.8 resulting cells that contain only the designed mutation (with no additional plasmid elements).

8.4.2.1 Making Competent Cells

We made competent cells using a lithium acetate (LiAc)/polyethylene glycol (PEG) based method [131] and frozen the cells at $-80°$ C in slow freezing container. Then for each transformation, we used the frozen competent cells and applied a LiAC/PEG/Single stranded DNA based method [131] to transform the cells. We made competent cells using BY4705 (MATα met15Δ0 ura3Δ0 trp1Δ63 leu2Δ0 his3Δ200 lys2Δ0 ade2Δ0) and used that to insert GFP. Then using the GFP labeled cells, we made competent cells and used those to implement the mutations.

8.4.2.2 Transformation and Homologous Recombination

As briefly explained above, we used a two step method to implement the GFP insertion (Figure 8.8) and mutations (Figure 8.9). In this method first the URA3 gene is used as a marker to select transformants, then the cells are transfered to FOA plates [132] so that URA3 is removed by a second homologous recombination and only the designed insertion/deletion/mutation remains. A key point that should not be overlooked is that as shown in Figures 8.8,8.9 depending on whether the homologous recombination occurs before or after the mutation point, the colony in the SD/-URA or FOA plate will have either the wild type or mutant sequence. The ratio of the two types of colonies depend on the length of the flanking homologous sequence assuming that homologous recombination occurs randomly along the homologous region. For this, in our design for mutations, we cloned 2 kb upstream and 1 kb downstream of

Table 8.4: **Efficiencies of the first and second steps of homologous recombination.** mutation positions are relative to 'ATG' start codon. The efficiency in -URA plate is defined as the ratio of type B colonies (see Figure 8.9)to all the colonies in SD-URA plate. The efficiency in FOA plate is defined as the ratio of type C colonies (see Figure 8.9) to all the colonies in FOA plate.

Mutation Position	step 1 efficiency (-URA plates)	step 2 efficiency (FOA plates)
0	0.30	0.68
−100	0.34	0.65
−200	0.38	0.61
−300	0.42	0.57
−400	0.45	0.54
−500	0.49	0.50
−600	0.53	0.46
−700	0.57	0.43
−800	0.60	0.39
−900	0.64	0.35
−1000	0.68	0.32

the HXT3 ATG start codon so that we get at least 30% efficiency for each of the steps, given any mutation within 1 kb upstream of the HXT3 start codon. Table (8.4) gives the expected efficiencies for each step for mutations at different positions at HXT3 promoter.

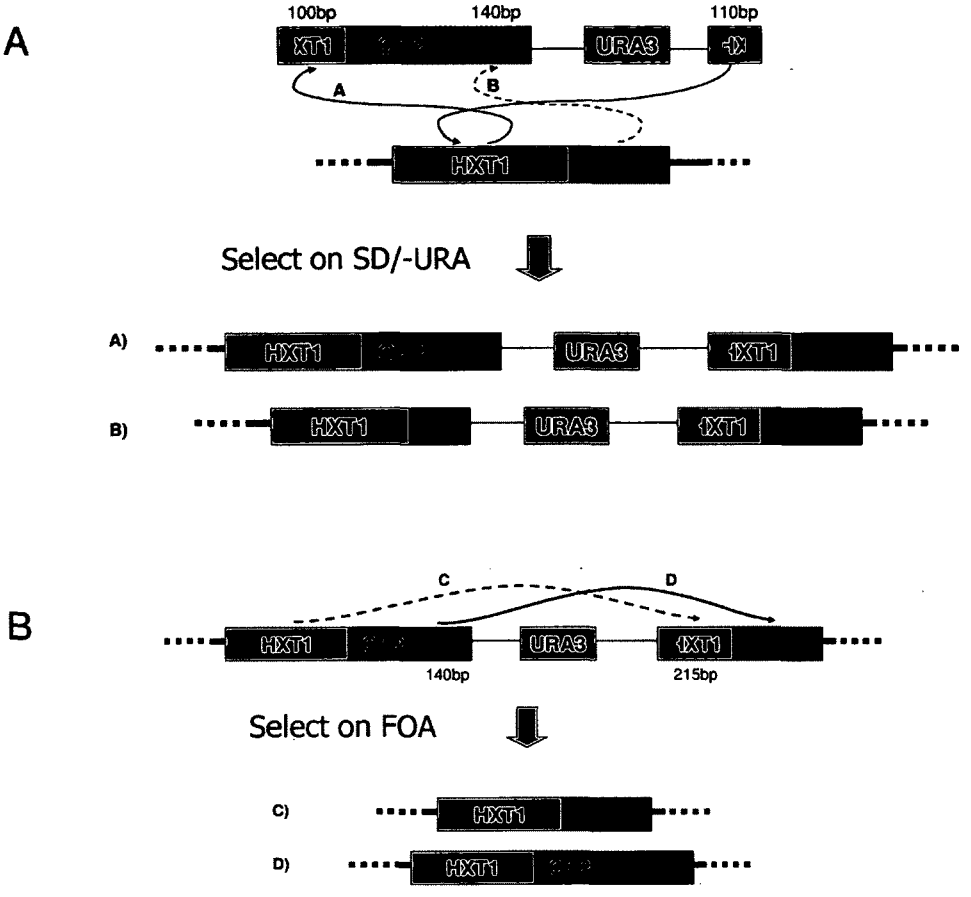


Figure 8.8: **Integration of GFP through homologous recombination** A) The linearized vector containing the cloned promoter region and URA3 gene is transformed to URA- strain. The linearized vector can hop in to the genome through homologous recombination as shown in the picture and the recombinants are selected on a -URA plate. Depending on whether the recombination occurs before or after the GFP insertion site, the colonies in -URA plate will be of either type A or B. Type B colonies are identified by fluorescence microscopy and/or sequencing and are transfered to FOA plates. B) Removal of URA3 gene through homologous recombination: Cells are grown on FOA plate to select against URA3 gene. Because of the homologous regions before and after the URA3 gene, URA3 gene can hop out of the genome through homologous recombination as shown. Depending on whether the recombination occurs before or after the mutation, the colonies in FOA plate will be of either type C or D. Type C colonies are identified by fluorescence microscopy and verified by sequencing.

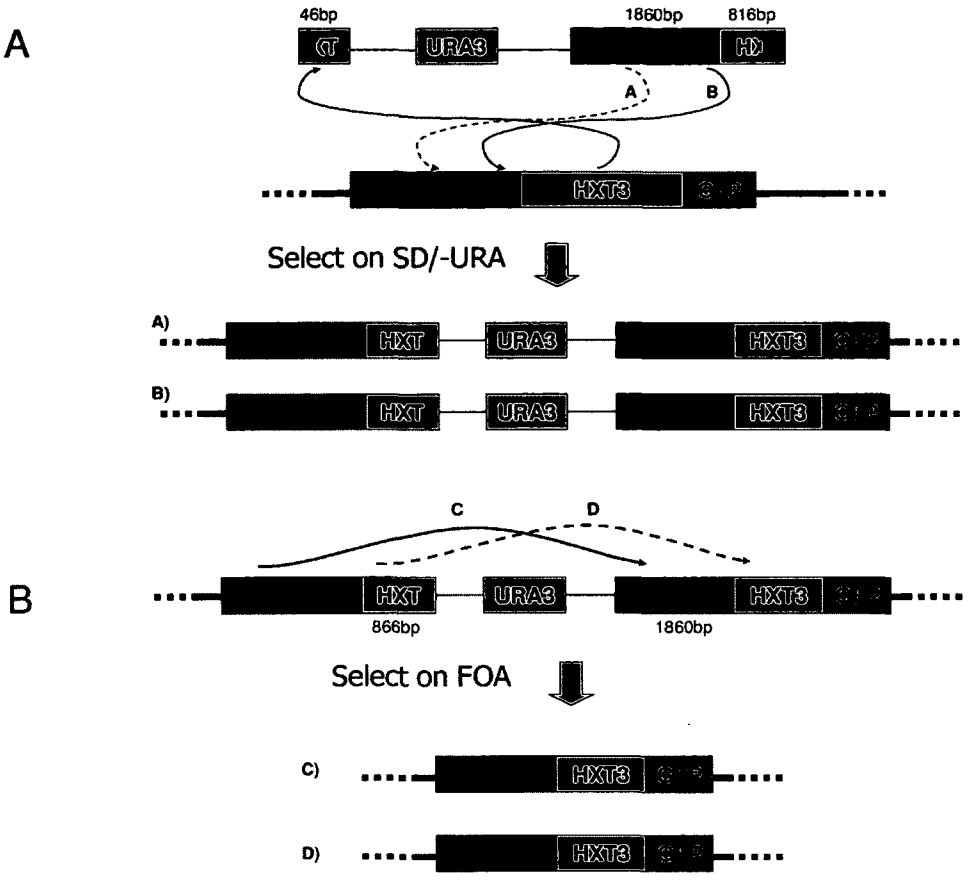


Figure 8.9: **Integration of foreign DNA to yeast genome through homologous recombination** A) The 'X' shows the mutation/insertion/deletion region in the HXT3 promoter. The linerized vector containing the cloned promoter region and URA3 gene is transformed to URA- strain. The linearized vector can hop in to the genome through homologous recombination as shown in the picture and the recombinants are selected on a -URA plate. Depending on whether the recombination occurs before or after the mutation, the colonies in -URA plate will be of either type A or B. Type B colonies are identified by restriction enzyme digestion or sequencing and transfered to FOA plates. B) Removal of URA3 gene through homologous recombination: Cells are grown on FOA plate to select against URA3 gene. Because of the homologous regions before and after the URA3 gene, URA3 gene can hop out of the genome through homologous recombination as shown. Depending on whether the recombination occurs before or after the mutation, the colonies in FOA plate will be of either type C or D. Type C colonies are identified by restriction enzyme digestion or sequencing and are verified by sequencing.

8.4.3 HXT3 Expression Measurement using Fluorescence Microscopy

To measure the HXT3 expression, we developed a method using fluorescence microscopy. In this method, we grew the cells on SD plates (we used SD plates for lower auto-fluorescence compared to YPD plates) with different concentrations of glucose. After two days, when the colonies are just big enough to be visible by naked eye, we image multiple colonies on each plate using a fluorescence microscope. Since different colonies in the plate have different sizes, and we take 2D images of 3D colonies, larger colonies will have higher average intensity as shown in Figure 8.10C . To estimate the expression level, we plot average intensity on the y axes against $\sqrt{\text{colony size}}$ on the x axes and fit a line, using the slope of the line as the measure for expression level. It is important to take images when the colonies are very small because when they grow big, they will consume the glucose in the plate and the local glucose level will change. Moreover in the above procedure as shown in Figure 8.10B we assume the colonies are round and the ratio of volume to the surface is proportional to the radius R, which may not hold for very large colonies. To avoid biases caused by different UV source and imager gains, we split each plate into two segments and streak a control strain with known expression level on one segment and the other strain (e.g. mutant) on the other segment on the same plate.

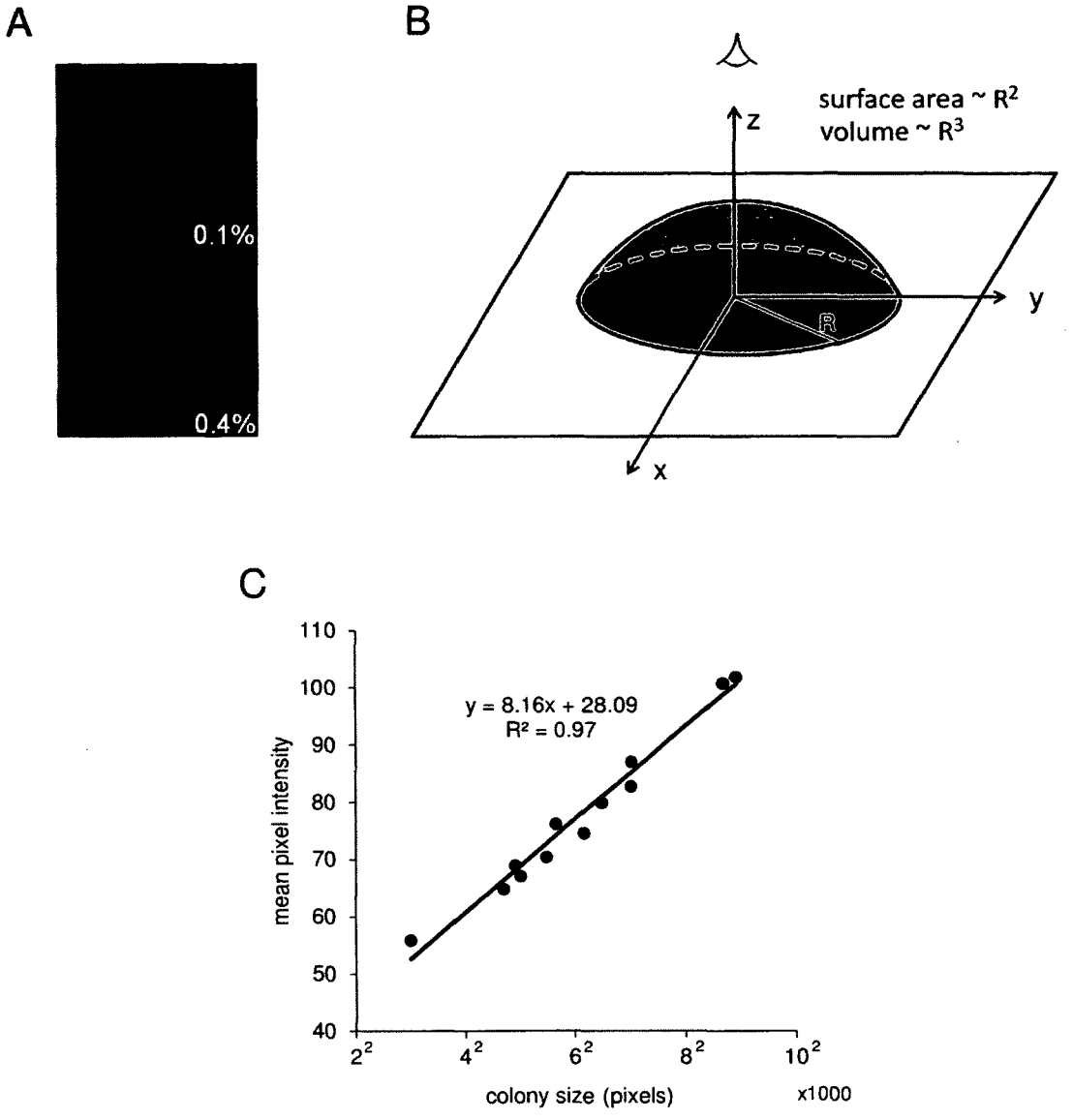


Figure 8.10: HXT3 expression measurement by fluorescence microscopy: In this method, we grow colonies on plates with different concentrations of glucose, and take images of the colonies. Then for each colony we estimate the colony size (surface area) and average GFP intensity from the image. A) Sample images of yeast colonies grown at 0.1% and 0.4% glucose concentrations. HXT3 is upregulated at high glucose concentration. B) 3D model for colony image. We use this model to relate the average pixel intensity to colony size in the 2D image. C) Mean pixel intensity from the colony image plotted against colony size for GFP labeled cell line. Gene expression is inferred from the slope of the regression line.

8.4.4 HXT3 Expression Measurement using Fluorescence Plate Reader

Using a fluorescence microscope, we can only measure the steady state expression level, since we have to wait for colonies to grow large enough so we can image them. Moreover, near the critical glucose concentration (0.2-0.4%) at which HXT expression level is very sensitive to small changes in glucose concentration, local variations in glucose concentration on the plate cause the colonies average intensity deviate from the linear model of Figure 8.10, resulting in lower precision of the method. We therefore used fluorescence plate reader to measure the time course response of the expression level after addition of glucose as described in the following. To measure the HXT3 expression response to addition of glucose for WT and mutated promoters, we made liquid culture, 5ml, SD/1%Gly+1%gal+0.1%glu inoculated from fresh colonies and incubated for 2 days in a 30° C shaker at 225 rpm. After 2 days cultures were saturated. We then spun down the cells and resuspended them in 10 ml fresh media (SD+1%Gly). Then aliquoted 230 μl to each well in a 96 well micro plate. Then added 20 μl mixture of glactose 20% and glucose 20%, so the combined final concentration of (gal+glu) is 1.6% for all the samples except for control well for which we added 20 μl of water. Glucose concentrations used were 0%, 0.1%, 0.2%, 0.4%, 0.8% and 1.6%. The 96 well plate was put in plate reader. Optical densities at 600 nm (OD600) and fluorescence (excitation 476nm, emission 515nm, cutoff 495 nm) were recorded every

15-20 minutes for 5 hours. The plate reader temperature was set at 30 degrees and plates were shaken for 2 seconds every 1-2 minutes, and 5 seconds before each read. We did three replicates on three different days. We used Loess [133] to interpolate the OD and GFP for at each minute. Then, to remove the background (autofluorescence), we subtracted the average signal for the two control wells from GFP measurement for each sample. Similarly we removed background for OD values. Then we used the ratio of the background corrected GFP and background corrected OD as the normalized GFP level.

8.5 Results

8.5.1 GFP Insertion

To test the feasibility of our microscopy method for measuring expression level, we first inserted yeast optimized gfp sequence [127] at ADH1 locus and obtained bright green colonies. We chose ADH1 because it was a highly expressed gene and ADH1 mutation was tested before in the literature [134, 135]. We then inserted GFP at the 3′ end of the HXT3 gene and obtained green colonies. The HXT3-GFP colonies grow about two times slower than wild type colonies. Moreover, we found that the HXT-GFP strain does not grow on 2% glactose plates with 0% glucose, however adding a very small amount of glucose (0.01%) to the media was enough to make the cells viable. This is interesting, because at very low glucose levels, HXT3 expression is

very low; however, it seems that in wild type strain, HXT3 is essential for growing on a media lacking glucose. The other possibility is that the HXT3-GFP fused protein is toxic for such growth condition. In our experiments we added the minimal amount of glucose (0.01%) to all the media. The growth rate is not sensitive to the amount of the minimal added glucose and cells grow at similar rate at (0.01% glucose + 2% glactose) compared to (0.10% glucose + 2% glactose) but does not grow on (0.00% glucose + 2% glactose).

We also tried to insert GFP in HXT1 locus using the same methodology, however we got no colony with GFP in SD/-URA plates. We also inserted GFP in HXT3 locus for the cell line which had already GFP inserted on ADH1 locus. This resulted in highly bright slow growing cells, as we expected, and we used this strain as control in our fluorescence experiments.

8.5.2 Rgta Binding Sites Deletion, Pal40 and M725

As a proof of concept and to test our ability to implement mutations we designed two double base pairs mutations: M725 (a mutation to convert 'CACCTGGAGGAG' → 'CACCCGGGGGAG' at 725bp upstream of HXT3 gene), and M680 (a mutation to convert 'CTTTTTCTTGAAAAAAG' → 'CTTTCTCTTGAAGAAAG' at 680 bp upstream of the HXT3 gene). We used the simple context-based model trained on tiling array data (*in vivo* data) to select mutations that make the promoter region more nucleosome bound or more nucleosome free. We used fusion PCR to implement these mutations. We

constructed the plasmids and verified them by sequencing, then linearized them and transformed them to GFP labeled cells. M725 mutation adds an XmaI restriction enzyme site so we could more easily check the colonies in SD/-URA plate and select the mutants. M680 however didn't have such a feature. Moreover the colonies on SD/-URA plates looked similar in terms of their sizes and fluorescence. Hence we only continued with M725 colonies. For the rest of the mutations, we always designed the mutation to add or remove a restriction enzyme site so we could easily verify the colonies by PCR amplification and RE digestion. We used the fluorescence microscopy method to measure the effect of M725 mutation. As shown in Figure 8.11, the mutation had only slightly changed the HXT3 expression response to glucose, however this change was in the direction that the model would expect. Interestingly, this double mutation that was the highest effecting double base pair mutation selected by the contex based model was converting the sequence at -725 bp upstream of HXT3 to a Reb1-like binding site (see Figure 3F in [45]) and is thought to be associated to widened nucleosome free regions in a significant number of genes [45].

We also deleted a region between 443 bp and 309 bp upstream of ATG (a total of 134 bp) which contained a cluster of 5 Rgta binding sites. The HXT3 promoter has 10 likely Rgta binding sites. We chose this region because the Rgta binding sites were in cluster. We used QuikChange Site-Directed Mutagenesis kit to implement this deletion. At this point we hadn't yet developed the algorithm described in section (8.4.1.3.1) to design the primers. The DNA sequencing results showed deletion of the

134 bp sequences as designed, but also insertion of 134 bp consisting of 3 times repeat of the flanking sequences used in primers. Fluorescence microscopy method showed that the Rgta binding sites deleted strain had partial level of HXT3 activity even at very low glucose level for which the wild type mutant had no detectable expression (Figure 8.11). This is consistent with our prediction that removing 5 out of 10 Rgta binding sites would lower the effect of this repressor protein. We also observed that the mutant had slightly lower HXT3 expression at high glucose. This may have been caused by deletion of some other binding sites within the 134bp region, specifically a putative RSC3/RSC30 binding sequence, or by reduction in Rgta function as an activator at high glucose [120].

We also deleted a 40bp region GAAAAAAAAATACTTTCTTTTTCTTGAAAAAAGAAAAAAA at 661 bp upstream of ATG. This sequence contained multiple poly A stretches and a semi-palindromic sequence. We used QuikChange Site-Directed Mutagenesis kit to implement this deletion. This time it worked fine. The DNA sequencing results showed deletion of this 40 bases had a small effect on the HXT3 expression response as shown in Figure 8.11, however this change was in the direction the our model had predicted.

Primers used for making these mutations are listed in Table (8.2).

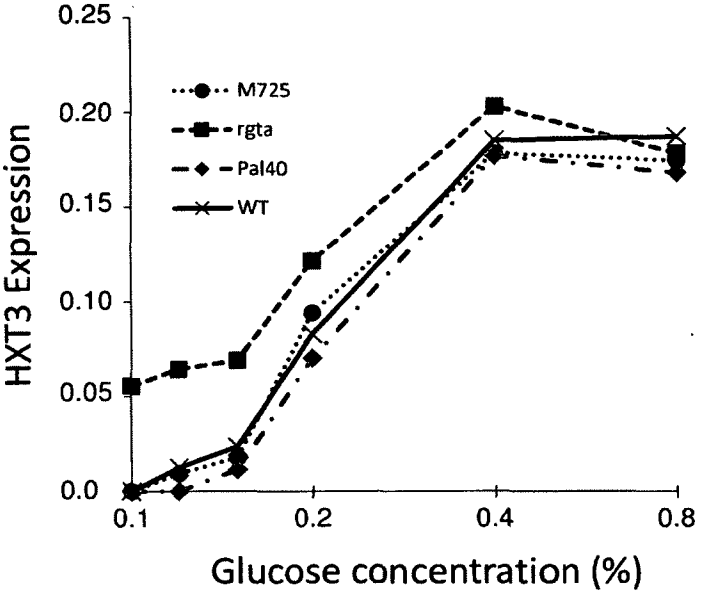


Figure 8.11: HXT3 expression measured at different glucose concentrations for cells with wild type promoter (WT), deletion of five out of ten Rgta binding sites (Rgta), deletion of 40bp of a nucleosome free region (Pal40) and a nucleosome destabilizing double base pairs substitution at 725bp upstream of the ATG start codon (M725)

8.5.3 PolyT Mutations, Selecting a Region

To choose a region for the HXT3 mutagenesis experiments, we first deleted 6-8 bp and inserted the T11 sequence ('TTTTTTTTTTT') in 210, 245, 560 and 600 bp upstream of ATG. We also made mutants with pairs of poly-T sequences inserted at (245,600), (560,600), (245,210), and (560,210) bp upstream of ATG start codon. Then we grow the cells on SD plates with different glucose concentration and used fluorescence microscope to evaluate the expression level. We found the highest effect for the double mutant with poly-T sequences inserted at (245,210), resulting in average GFP signal intensity to increase about 66% from 41(7) to 68(10). Figure 8.12 shows the average HXT3 response at 0.12% glucose. Interestingly, insertion of poly-T in either -245 or -210 bp upstream of HXT3 didn't have a detectable effect compared to wild type promoter; however when we added poly-T at both -245 and -210, the expression level significantly increased. This is consistent with the model prediction that destabilizing the nucleosome at the HXT3 promoter would allow increased HXT3 expression at low glucose concentrations.

8.5.4 Random Mutants

Based on the poly-T insertion experiments, we chose the (-245,-210) region upstream of HXT3 and implemented 7 mutations selected from a pool of random mutations as described in 8.3.1. Table 8.5 gives the list of these mutations. We measured

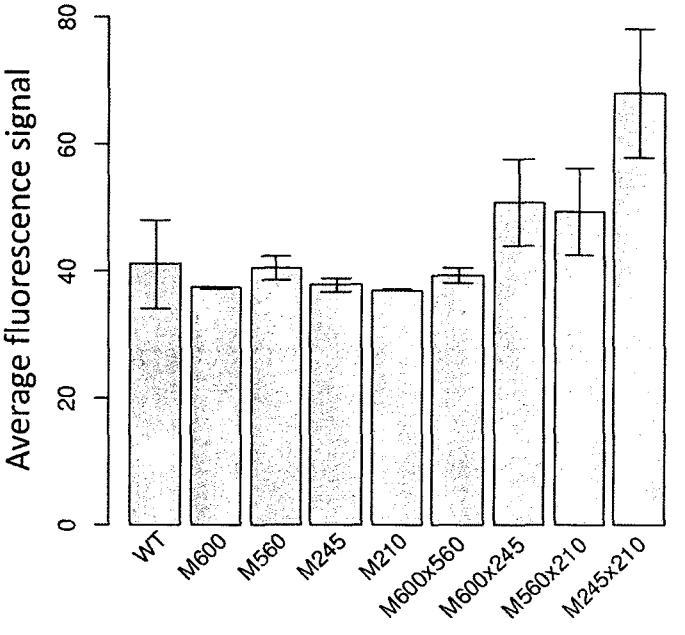


Figure 8.12: Comparison of the effect of inserting poly-T sequence at different distances from ATG start codon. The x-axis shows the strain name, WT is the wild type promoter, M600, M560, M245, and M210 are mutants with insertion of one poly-T sequence at 600, 560, 245, and 210 bp upstream of HXT3 gene start codon, and M600x560, M600x245, M560x210, and M245x210 are double mutants containing two inserted poly-T sequences. The y-axis is average green intensity of colonies grown on SD plates with 0.12% glucose and 1.88% glactose. The error bars show the standard deviation among different colonies.

Table 8.5: **List of mutations at 245 and 210 bp upstream of HXT3**

Name	Sequence ($5' \rightarrow 3'$)
mutantA	CAAATATCTTCTAAGCTTGGAGAAAGAAACAACTCAATAACGATATCCCTCAGCAATGTTG
mutantB	CAAAAAAAATGTACTAGTGGAGAAAGAAACAACTCAATAACGATAAGCTTTTTTTTTTTTG
mutantC	CAAAAGCTTTTTTTTTTTGGAGAAAGAAACAACTCAATAACGATATCCCACTTTTTTGTTG
mutantD	CAAAAGCTTTTTTTTTTTGGAGAAAGAAACAACTCAATAACGATCACAAAAAAAAAAGCTTG
mutantE	CAAAAGCTTTTTTTTTTTGGAGAAAGAAACAACTCAATAACGATATCAAAAAAATTTTTTG
mutantF	CAATGGACCACGCTAGCTGGAGAAAGAAACAACTCAATAACGATGGATATCCACCACCTTG
mutantG	CAAGACCATCTAGATGTTGGAGAAAGAAACAACTCAATAACGATTGGATCTGGATCCATTG
WT	CAACGACCTTCTGGAGAAAGAAACAACTCAATAACGATGTGGGACATTG
M210	CAACGACCTTCTGGAGAAAGAAACAACTCAATAACGATAAAAAAAAAAAGCTTG
M245	CAAAAAAAAAAAACTAGTGGAGAAAGAAACAACTCAATAACGATGTGGGACATTG
M245x210	CAAAAAAAAAAAACTAGTGGAGAAAGAAACAACTCAATAACGATAAAAAAAAAAAGCTTG

the HXT3 response to addition of different concentrations of glucose (0%, 0.1%, 0.2%, 0.4%, 0.8%, 1.6%) over 5 hours using fluorescence plate reader. We also used GSVM (g), Phase-based context-based (ph), and simple context-based (cb) models trained on *in vivo* (ypd) and *in vitro* (ivtr) data using all the chromosomes excluding chromosome 4, to score different mutants. We also used simple context-based model with context size (n=1) (cbn1) for comparison. Figure 8.13 shows predicted nucleosome occupancy for different mutants using different models. These scores are normalized (shifted and scaled) to match the mean and variance of givtr model scores. It can be observed that there is a high correlation between the predicted scores using different models. The correlations ranges from 0.949 between cbn1ivtr and phypd, to 0.999 for cbn1ivtr and cbn1ypd. Then we used the experimental data of HXT3 expression

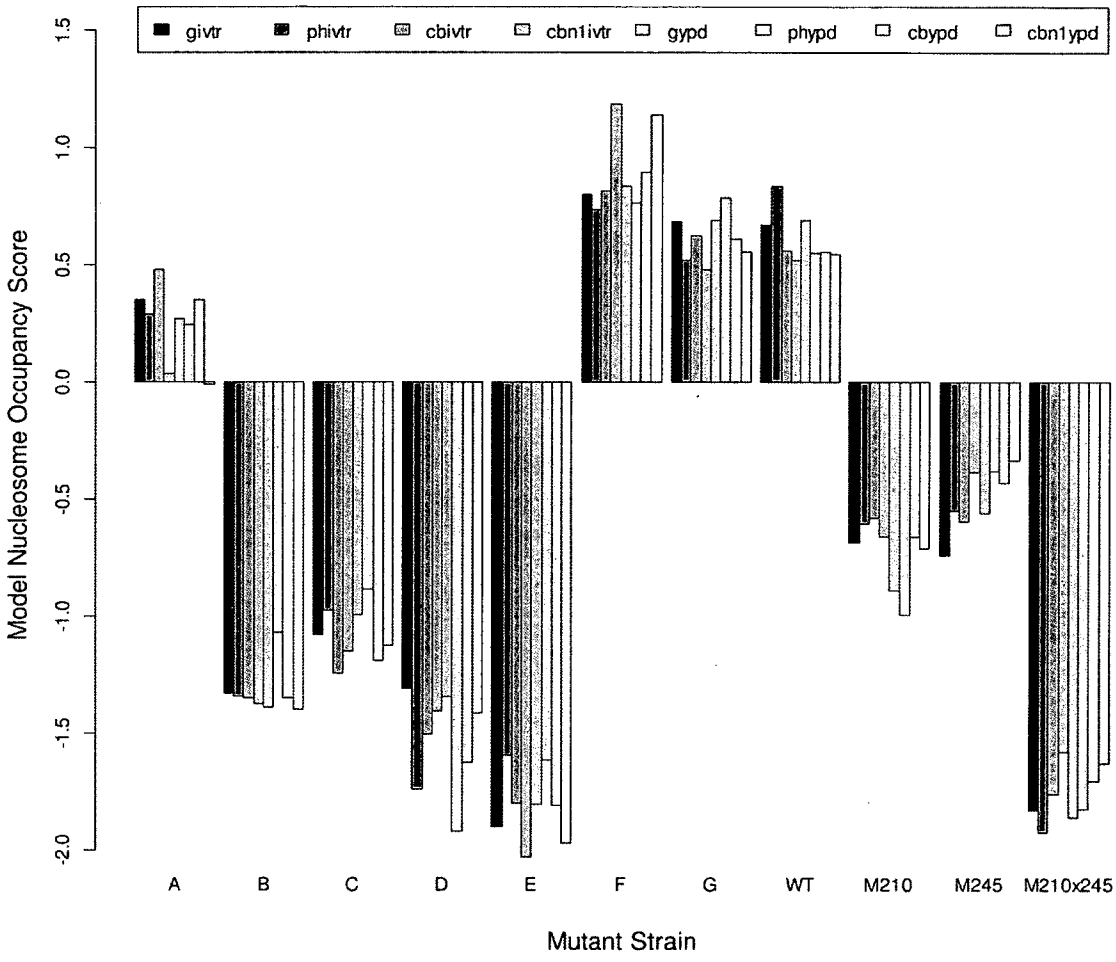


Figure 8.13: Nucleosome occupancy score for different mutants using different models

Table 8.6: **Nucleosome occupancy score and HXT3 expression correlation:** Pearson correlation coefficients of experimental data with different nucleosome positioning models trained using *in vivo* and *in vitro* data.

	GSVM	CB+Phase	CB(n=12)	CB(n=1)
in vitro	−0.67	−0.64	−0.67	−0.64
in vivo	−0.66	−0.61	−0.65	−0.65

to compare with model scores. Figure 8.14 shows the normalized GFP for different concentrations of added glucose. To compare the effect of mutations on HXT3 regulation, we calculated the average of the normalized GFP level from time 150 min to 180 min after addition of glucose for WT promoter and each of the mutant strains. Figure 8.16 shows the average expression against nucleosome occupancy score predicted by GSVM model train on *in vitro* data. As it can be observed, there is a negative correlation (Pearson correlation coefficient = −0.67) between nucleosome occupancy score and expression, so in general, mutations that caused the region to become more nucleosome free have resulted in upregulation of HXT3 after addition of 0.1% glucose consistent with our hypothesis that nucleosome removal facilitates the upregulation of HXT3 at higher glucose concentrations. The Pearson correlation coefficient of the experimental data with other sequence based models are given in Table (8.6). It can be observed that GSVM and simple context-based model trained using *in vitro* data achieved the highest negative correlation. Compared to *in vitro* simple context-based model with context length (n=1), GSVM-invitro and Simple

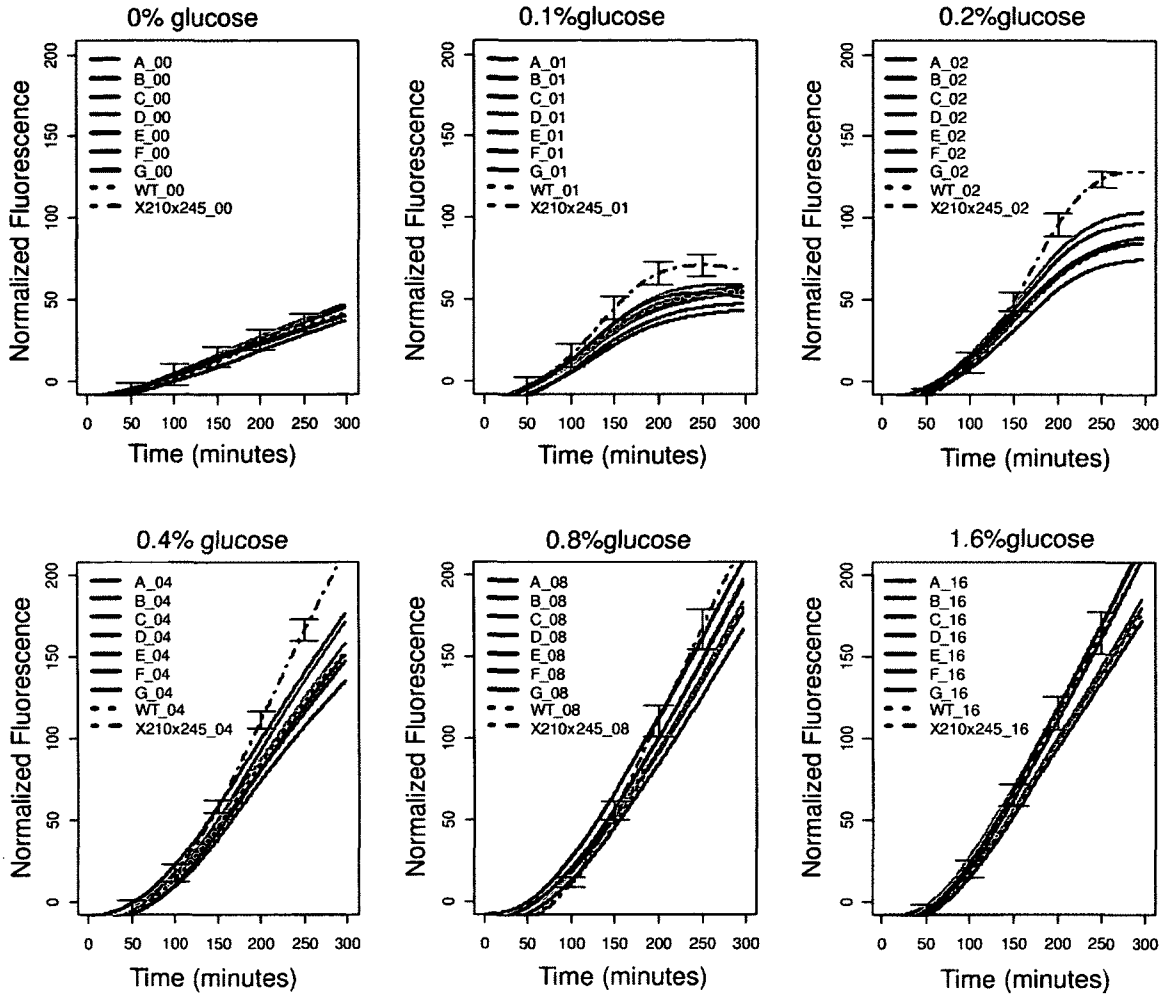


Figure 8.14: HXT3 response to addition of different amounts of glucose for different strains

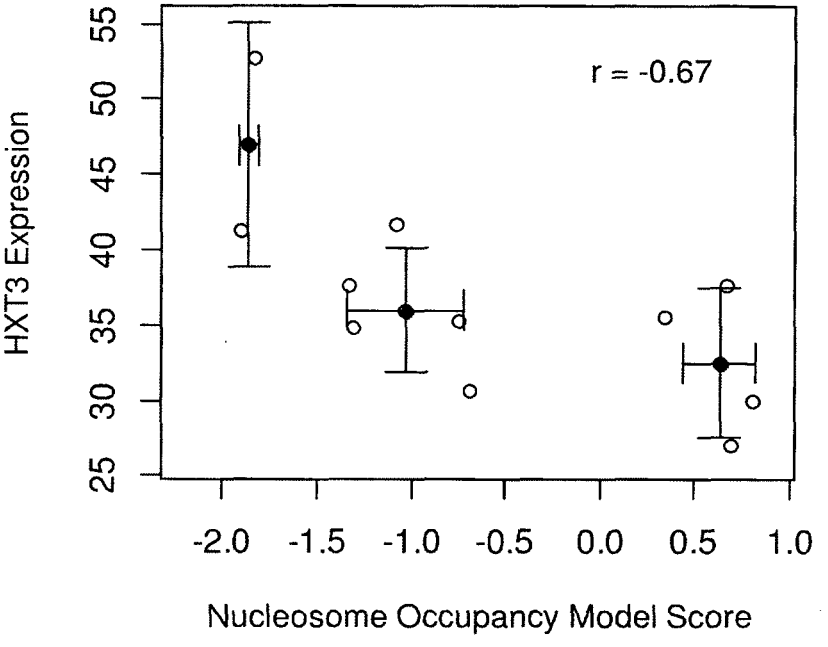


Figure 8.15: Comparison of experimental HXT3 expression and sequence-based model *in vitro* nucleosome affinity score

context-based model (n=12) gave 0.03(0.02), and 0.02(0.01) higher negative correlation coefficient respectively, where the numbers in parentheses are standard deviation over the three replicates.

8.6 Summary and Discussion

In this chapter we described an experimental system to validate the nucleosome positioning models developed in this thesis. We made mutations in the HXT3 promoter at a region that undergoes nucleosome remodeling upon glucose addition, then measured the HXT3 expression over time at different glucose concentrations. Our results show a significant negative correlation between model predicted nucleosome occupancy and expression level suggesting that nucleosome remodeling at HXT3 promoter has a functional role in glucose-dependent regulation of HXT3. This is a novel finding as the previous models for HXT3 were only based on Rgta. Figure 8.16 shows a possible model for HXT3 regulation. This model is consistent with our results. It is also consistent with ChIP results that suggest RSC interacts with HXT3 promoter [126, 136]. One mechanism that Rgta may inhibit RSC could be through competitive binding to DNA. Some of Rgta binding sites overlap with CGCG motif which is a known binding sequence for RSC3 [114]. So we hypothesize that Rgta binding, inhibits RSC complex at low glucose. At high glucose, Rgta is phosphorylated and detached from the promoter, allowing RSC complex to interact with the

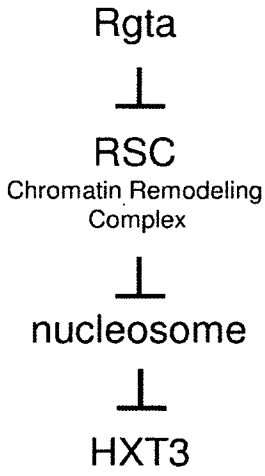


Figure 8.16: HXT3 regulation model: This figure shows a possible model for HXT3 regulation which is consistent with our data and previous studies. At low glucose, Rgta is bound to its binding sites on HXT3 promoter. Binding of Rgta inhibits RSC complex from binding to the promoter. At high glucose, Rgta is phosphorylated and detached from the promoter. This allows the RSC complex to interact with the promoter and to remove the nucleosome at the promoter. This facilitates interactions of other TFs and the transcription machinery resulting in up-regulation of the gene.

promoter and to remove the nucleosome. This facilitates interactions of other TFs and the transcription machinery resulting in up-regulation of the gene.

The experimental system that was proposed in this chapter can be further improved by measuring (or controlling) glucose level at different time points along the experiment, and also by quantifying the cell concentrations at different time points based on the OD measurements, to obtain a better picture of the effect of the mutations on the global response of the gene to different glucose concentrations. Also an improvement to the current protocol is to start with cells growing at log-phase to avoid possible interferences originated from cells in saturation. Finally, more mutagenesis experiments should be designed to test and validate the model for HXT3 regulation and to elucidate the relation between Rgta, RSC complex and nucleosome occupancy.

Chapter 9

General Discussions

9.1 Summary and Discussion

In this thesis we presented a novel non-parametric normalization procedure (group normalization) for microarray data that integrates the array wide normalization and probe effect correction into one simple step relying on the fact that a typical microarray has such a large number of probes that we can estimate the sequence biases of hybridization (the dynamic parameters of the probes) from the response of a group of similar probes. We also described an approach based on this technique which highlights regions of significant signal changes between two experimental conditions (cross normalization). We used this method to normalize and compare genome wide nucleosome positioning data in yeast at different glucose concentrations conditions.

We also developed sequence based models for nucleosome affinity. The phase-

dependent context-based model inspired by the geometry of the nucleosome and the observation that there is a ~10 bp periodicity in some dinicleotide frequencies along the nucleosome sequence, significantly improved prediction of nucleosome affinities for *in vivo* and *in vitro* datasets. We also developed a novel method for robust estimation of k-mer frequencies using gapped k-mers and developed a novel sequence similarity score that we used as a SVM kernel to model nucleosome affinities. This sequence kernel is very general and could be applied to different classification problems. We showed its applicability for modeling nucleosome positioning data, mammalian enhancer data, and CTCF binding site modeling. We also developed a novel data structure and fast algorithm for the kernel calculation which makes this approach feasible for large genome-wide datasets.

We also investigated the average nucleosome positioning patterns near transcription start sites (TSSs) in yeast. By proposing a simple reconstruction method, we showed that the asymmetric average nucleosome pattern around TSSs is explainable by the asymmetric distribution of Reb1, Abf1, and Rap1 binding sites around TSSs highlighting the role of ATP-dependent factors for the formation of the average regular pattern. We also showed that the average pattern is not representative of individual promoters and different classes of promoters (linked to different gene classes) have different nucleosome organizations that may be established through different mechanisms.

Finally we developed an experimental system using glucose-dependent yeast HXT3

gene to test the sequence based models predictions. Our results show that the change in the intrinsic DNA/histone affinity in the promoter nucleosome of HXT3 can affect gene expression and its response to glucose, suggesting that the magnitude of the intrinsic binding energies are comparable to other factors (e.g. ATP-dependent factors) that are affecting the nucleosome positioning in HXT3 promoter. We observed a statistically significant negative correlation of (-0.67) between the model predicted nucleosome affinity and HXT3 expression. This shows that the sequence based models for nucleosome positioning can be used to improve our predictive models for *in vivo* gene regulation.

9.2 Future Directions

(i) In chapter 3 we proposed a normalization procedure for genomic data. We use the reference set probes to find an estimation for the normalized signal of each probe. A possible extension of this model would incorporate an estimate of the variance of the probe signal from the reference set, and use that as a measure of the reliability of a probe. Then we can use the variance to weigh different probe values when combining multiple probe values similar to the approach used in MAT algorithm.

(ii) As was shown in this thesis, different mechanisms effect nucleosome positioning *in vivo*. Here we proposed improved sequence based models for intrinsic DNA/histone interactions. An extension to this work would be to build a combined model to in-

tegrate the effect of ATP-dependent remodeling factors and intrinsic DNA/histone interactions by quantifying the effect of ATP-dependent factors and the intrinsic binding energies. A simple approach inspired by the average pattern reconstruction algorithm to quantify the relative effect of ATP-dependent factors and intrinsic interactions is to decompose the nucleosome occupancy signal around TSS to ATP-dependent and intrinsic interactions. Results from the reconstruction algorithm in chapter 7 showed that even a simple additive model could be useful to give a first order approximation of the contribution of each mechanism.

(iii) The sequence similarity score and kernel approximation algorithm showed very promising results on nucleosome positioning and mammalian enhancer sequence modeling and there is room to improve them. One way to improve this similarity score algorithm is to initialize the weights for different number of mismatches based on a given value of ℓ and k and then use a numerical optimization algorithm such as steepest descent or conjugate gradient to optimize the weights to obtain higher classification performances.

(iv) The experimental system that was proposed can be further improved by measuring and controlling glucose level at different time points along the experiment, and also by quantifying the cell concentrations at different time points based on the OD measurements, to obtain a better picture of the effect of the mutations on the global response of the gene to different glucose concentrations.

(v) Finally, more mutagenesis experiments should be designed to test and vali-

date the model for HXT3 regulation and to elucidate the relation between Rgta, RSC complex and nucleosome occupancy. Also more direct measurement of nucleosome occupancy at HXT3 promoter for the mutant and wild type strains by MNase digestion experiments followed by nucleosomal DNA purification and qPCR can further improve this work.

Bibliography

[1] M. A. Beer and S. Tavazoie, "Predicting gene expression from sequence," *Cell*, vol. 117, no. 2, pp. 185–198, Apr. 2004, PMID: 15084257. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15084257

[2] D. Voet, *D. Voet's Biochemistry 3rd (Third) edition (Biochemistry [Hardcover]*, 3rd ed. Wiley, Jan. 2004.

[3] K. Luger, A. W. M|[auml]|der, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8|[thinsp]||[Aring]| resolution," *Nature*, vol. 389, no. 6648, pp. 251–260, Sep. 1997. [Online]. Available: http://www.nature.com/nature/journal/v389/n6648/full/389251a0.html

[4] L. A. Boyer, X. Shao, R. H. Ebright, and C. L. Peterson, "Roles of the histone H2A-H2B dimers and the (H3-H4)2Tetramer in nucleosome remodeling by the SWI-SNF complex," *Journal of Biological Chemistry*,

vol. 275, no. 16, pp. 11 545–11 552, Apr. 2000. [Online]. Available: http://www.jbc.org/content/275/16/11545

[5] Y. Park, J. V. Chodaparambil, Y. Bao, S. J. McBryant, and K. Luger, "Nucleosome assembly protein 1 exchanges histone H2A-H2B dimers and assists nucleosome sliding," *Journal of Biological Chemistry*, vol. 280, no. 3, pp. 1817–1825, Jan. 2005. [Online]. Available: http://www.jbc.org/content/280/3/1817

[6] F. H. Lam, D. J. Steger, and E. K. O'Shea, "Chromatin decouples promoter threshold from dynamic range," *Nature*, vol. 453, no. 7192, pp. 246–250, May 2008, PMID: 18418379. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18418379

[7] J. C. Lin, S. Jeong, G. Liang, D. Takai, M. Fatemi, Y. C. Tsai, G. Egger, E. N. Gal-Yam, and P. A. Jones, "Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island," *Cancer Cell*, vol. 12, no. 5, pp. 432–444, Nov. 2007, PMID: 17996647. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17996647

[8] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thström, Y. Field, I. K. Moore, J. Z. Wang, and J. Widom, "A genomic code for nucleosome positioning," *Nature*, vol. 442, no. 7104, pp. 772–778, Aug. 2006, PMID: 16862119. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16862119

[9] J. M. Moreira and S. Holmberg, "Nucleosome structure of the yeast CHA1

promoter: analysis of activation-dependent chromatin remodeling of an RNA-polymerase-II-transcribed gene in TBP and RNA pol II mutants defective in vivo in response to acidic activators." *The EMBO Journal*, vol. 17, no. 20, pp. 6028–6038, Oct. 1998, PMID: 9774346 PMCID: PMC1170929.

[10] W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow, "A high-resolution atlas of nucleosome occupancy in yeast," *Nature Genetics*, vol. 39, no. 10, pp. 1235–1244, Oct. 2007, PMID: 17873876. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17873876

[11] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, "The DNA-encoded nucleosome organization of a eukaryotic genome," *Nature*, vol. 458, no. 7236, pp. 362–366, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1038/nature07667

[12] S. Shivaswamy, A. Bhinge, Y. Zhao, S. Jones, M. Hirst, and V. R. Iyer, "Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation," *PLoS Biology*, vol. 6, no. 3, p. e65, Mar. 2008, PMID: 18351804. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18351804

[13] Q. He, C. Yu, and R. H. Morse, "Dispersed mutations in histone h3 that affect transcriptional repression and chromatin structure of the CHA1 promoter in

saccharomyces cerevisiae," *Eukaryotic Cell*, vol. 7, no. 10, pp. 1649–1660, Oct. 2008. [Online]. Available: http://ec.asm.org/cgi/content/abstract/7/10/1649

[14] Y. Zhang, Z. Moqtaderi, B. P. Rattner, G. Euskirchen, M. Snyder, J. T. Kadonaga, X. S. Liu, and K. Struhl, "Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo," *Nat Struct Mol Biol*, vol. 16, no. 8, pp. 847–852, 2009. [Online]. Available: http://dx.doi.org/10.1038/nsmb.1636

[15] T. N. Mavrich, C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, D. S. Gilmour, I. Albert, and B. F. Pugh, "Nucleosome organization in the drosophila genome," *Nature*, vol. 453, no. 7193, pp. 358–362, May 2008. [Online]. Available: http://dx.doi.org/10.1038/nature06929

[16] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, and S. M. Johnson, "A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning," *Genome Research*, vol. 18, no. 7, pp. 1051–1063, Jul. 2008, PMID: 18477713. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18477713

[17] S. Sasaki, C. C. Mello, A. Shimada, Y. Nakatani, S. Hashimoto, M. Ogawa, K. Matsushima, S. G. Gu, M. Kasahara, B. Ahsan, A. Sasaki, T. Saito,

Y. Suzuki, S. Sugano, Y. Kohara, H. Takeda, A. Fire, and S. Morishita, "Chromatin-Associated periodicity in genetic variation downstream of transcriptional start sites," *Science*, vol. 323, no. 5912, pp. 401–404, Jan. 2009. [Online]. Available: http://www.sciencemag.org/content/323/5912/401

[18] D. E. Schones, K. Cui, S. Cuddapah, T. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao, "Dynamic regulation of nucleosome positioning in the human genome," *Cell*, vol. 132, no. 5, pp. 887–898, Mar. 2008, PMID: 18329373. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18329373

[19] A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, and A. Sidow, "Determinants of nucleosome organization in primary human cells," *Nature*, vol. 474, no. 7352, pp. 516–520, Jun. 2011, PMID: 21602827. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21602827

[20] F. Ozsolak, J. S. Song, X. S. Liu, and D. E. Fisher, "High-throughput mapping of the chromatin structure of human promoters," *Nature Biotechnology*, vol. 25, no. 2, pp. 244–248, Jan. 2007. [Online]. Available: http://www.nature.com/nbt/journal/v25/n2/full/nbt1279.html

[21] H. Chung, I. Dunkel, F. Heise, C. Linke, S. Krobitsch, A. E. Ehrenhofer-Murray, S. R. Sperling, and M. Vingron, "The effect of micrococcal nuclease digestion on nucleosome positioning data," *PLoS ONE*, vol. 5, no. 12, p. e15754, Dec. 2010. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0015754

[22] M. Visnapuu and E. C. Greene, "Single-molecule imaging of DNA curtains reveals intrinsic energy landscapes for nucleosome deposition," *Nature Structural & Molecular Biology*, vol. 16, no. 10, pp. 1056–1062, Sep. 2009. [Online]. Available: http://www.nature.com/nsmb/journal/v16/n10/full/nsmb.1655.html

[23] D. Tolkunov, A. V. Morozov, and A. McPherson, "Genomic studies and computational predictions of nucleosome positions and formation energies." Academic Press, 2010, vol. Volume 79, pp. 1–57. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1876162310790015

[24] D. Tillo and T. R. Hughes, "G+C content dominates intrinsic nucleosome occupancy," *BMC Bioinformatics*, vol. 10, no. 1, p. 442, Dec. 2009. [Online]. Available: http://www.biomedcentral.com/1471-2105/10/442

[25] I. Whitehouse, O. J. Rando, J. Delrow, and T. Tsukiyama, "Chromatin remodelling at promoters suppresses antisense transcription," *Nature*, vol. 450, no. 7172, pp. 1031–1035, Dec. 2007, PMID: 18075583. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18075583

[26] O. Bell, V. K. Tiwari, N. H. Thomä, and D. Schbeler, "Determinants and dynamics of genome accessibility," *Nature Reviews Genetics*, vol. 12, no. 8, pp. 554–564, Jul. 2011. [Online]. Available: http://www.nature.com/nrg/journal/v12/n8/full/nrg3017.html

[27] R. Padinhateeri and J. F. Marko, "Nucleosome positioning in a model of active chromatin remodeling enzymes," *Proceedings of the National Academy of Sciences*, vol. 108, no. 19, pp. 7799 –7803, May 2011. [Online]. Available: http://www.pnas.org/content/108/19/7799.abstract

[28] T. J. Sarnowski, G. Ros, J. Jsik, S. Swiezewski, S. Kaczanowski, Y. Li, A. Kwiatkowska, K. Pawlikowska, M. Kozbial, P. Kozbial, C. Koncz, and A. Jerzmanowski, "SWI3 subunits of putative SWI/SNF Chromatin-Remodeling complexes play distinct roles during arabidopsis development," *The Plant Cell Online*, vol. 17, no. 9, pp. 2454 –2472, 2005. [Online]. Available: http://www.plantcell.org/content/17/9/2454.abstract

[29] J. J. F. A. van Vugt, M. de Jager, M. Murawska, A. Brehm, J. van Noort, and C. Logie, "Multiple aspects of ATP-Dependent nucleosome translocation by RSC and mi-2 are directed by the underlying DNA sequence," *PLoS ONE*, vol. 4, no. 7, p. e6345, Jul. 2009. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0006345

[30] I. Whitehouse and T. Tsukiyama, "Antagonistic forces that position nucleosomes in vivo," *Nature Structural & Molecular Biology*, vol. 13, no. 7, pp. 633–640, Jul. 2006. [Online]. Available: http://www.nature.com/nsmb/journal/v13/n7/full/nsmb1111.html

[31] Y. Fu, M. Sinha, C. L. Peterson, and Z. Weng, "The insulator binding pro-

tein CTCF positions 20 nucleosomes around its binding sites across the human genome," *PLoS Genetics*, vol. 4, no. 7, Jul. 2008, PMID: 18654629 PMCID: 2453330.

[32] T. Wasson and A. J. Hartemink, "An ensemble model of competitive Multi-Factor binding of the genome," *Genome Research*, vol. 19, no. 11, pp. 2101–2112, Nov. 2009. [Online]. Available: http://genome.cshlp.org/content/19/11/2101

[33] L. A. Mirny, "Nucleosome-Mediated cooperativity between transcription factors," *Proceedings of the National Academy of Sciences*, vol. 107, no. 52, pp. 22 534–22 539, Dec. 2010. [Online]. Available: http://www.pnas.org/content/107/52/22534

[34] T. Raveh-Sadka, M. Levo, and E. Segal, "Incorporating nucleosomes into thermodynamic models of transcription regulation," *Genome Research*, vol. 19, no. 8, pp. 1480–1496, Aug. 2009. [Online]. Available: http://genome.cshlp.org/content/19/8/1480

[35] V. B. Teif and K. Rippe, "Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities," *Nucleic Acids Research*, vol. 37, no. 17, pp. 5641–5655, Sep. 2009, PMID: 19625488. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19625488

[36] A. Weiner, A. Hughes, M. Yassour, O. J. Rando, and N. Friedman, "High-

resolution nucleosome mapping reveals transcription-dependent promoter packaging," *Genome Research*, vol. 20, no. 1, pp. 90–100, Jan. 2010, PMID: 19846608. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19846608

[37] A. Thström, L. M. Bingham, and J. Widom, "Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning," *Journal of molecular biology*, vol. 338, no. 4, pp. 695–709, May 2004, PMID: 15099738. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15099738

[38] G. Yuan and J. S. Liu, "Genomic sequence is highly predictive of local nucleosome depletion," *PLoS Computational Biology*, vol. 4, no. 1, p. e13, Jan. 2008, PMID: 18225943. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18225943

[39] H. E. Peckham, R. E. Thurman, Y. Fu, J. A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng, "Nucleosome positioning signals in genomic DNA," *Genome Research*, vol. 17, no. 8, pp. 1170–1177, Aug. 2007, PMID: 17620451. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17620451

[40] R. Kornberg, "The location of nucleosomes in chromatin: specific or statistical?" *Nature*, vol. 292, no. 5824, pp. 579–580, 1981. [Online]. Available: http://dx.doi.org/10.1038/292579a0

[41] R. D. Kornberg and L. Stryer, "Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism," *Nucleic Acids Research*, vol. 16, no. 14A, pp. 6677–6690, Jul. 1988, PMID: 3399412. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3399412

[42] T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, S. C. Schuster, I. Albert, and B. F. Pugh, "A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome," *Genome Research*, vol. 18, no. 7, pp. 1073–1083, Jul. 2008, PMID: 18550805. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18550805

[43] I. Albert, T. N. Mavrich, L. P. Tomsho, J. Qi, S. J. Zanton, S. C. Schuster, and B. F. Pugh, "Translational and rotational settings of H2A.Z nucleosomes across the saccharomyces cerevisiae genome," *Nature*, vol. 446, no. 7135, pp. 572–576, Mar. 2007, PMID: 17392789. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17392789

[44] Y. Zhang, Z. Moqtaderi, B. P. Rattner, G. Euskirchen, M. Snyder, J. T. Kadonaga, X. S. Liu, and K. Struhl, "Evidence against a genomic code for nucleosome positioning reply to [ldquo]Nucleosome sequence preferences influence in vivo nucleosome organization[rdquo]," *Nat Struct Mol Biol*, vol. 17, no. 8, pp. 920–923, 2010. [Online]. Available: http://dx.doi.org/10.1038/nsmb0810-920

[45] P. D. Hartley and H. D. Madhani, "Mechanisms that specify promoter nucleosome location and identity," *Cell*, vol. 137, no. 3, pp. 445–458, May 2009, PMID: 19410542. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19410542

[46] H. Zhang, D. N. Roberts, and B. R. Cairns, "Genome-wide dynamics of htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss," *Cell*, vol. 123, no. 2, pp. 219–231, Oct. 2005, PMID: 16239141. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16239141

[47] S. J. Zanton and B. F. Pugh, "Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock," *Genes & development*, vol. 20, no. 16, pp. 2250–2265, Aug. 2006, PMID: 16912275. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16912275

[48] B. E. Bernstein, C. L. Liu, E. L. Humphrey, E. O. Perlstein, and S. L. Schreiber, "Global nucleosome occupancy in yeast," *Genome biology*, vol. 5, no. 9, p. R62, 2004, PMID: 15345046. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15345046

[49] R. M. Raisner, P. D. Hartley, M. D. Meneghini, M. Z. Bao, C. L. Liu, S. L. Schreiber, O. J. Rando, and H. D. Madhani, "Histone variant H2A.Z

marks the 5' ends of both active and inactive genes in euchromatin," *Cell*, vol. 123, no. 2, pp. 233–248, Oct. 2005, PMID: 16239142. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16239142

[50] I. Whitehouse, A. Flaus, B. R. Cairns, M. F. White, J. L. Workman, and T. Owen-Hughes, "Nucleosome mobilization catalysed by the yeast SWI/SNF complex," *Nature*, vol. 400, no. 6746, pp. 784–787, Aug. 1999, PMID: 10466730. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10466730

[51] B. G. Wilson and C. W. M. Roberts, "SWI/SNF nucleosome remodellers and cancer," *Nature reviews. Cancer*, vol. 11, no. 7, pp. 481–492, Jul. 2011, PMID: 21654818. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21654818

[52] A. Hamiche, R. Sandaltzopoulos, D. A. Gdula, and C. Wu, "ATP-dependent histone octamer sliding mediated by the chromatin remodeling complex NURF," *Cell*, vol. 97, no. 7, pp. 833–842, Jun. 1999, PMID: 10399912. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10399912

[53] Z. Zhang, C. J. Wippo, M. Wal, E. Ward, P. Korber, and B. F. Pugh, "A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome," *Science*, vol. 332, no. 6032, pp. 977–980, May 2011. [Online]. Available: http://www.sciencemag.org/content/332/6032/977

[54] K. A. Zawadzki, A. V. Morozov, and J. R. Broach, "Chromatin-

dependent transcription factor accessibility rather than nucleosome remodeling predominates during global transcriptional restructuring in saccharomyces cerevisiae," *Molecular Biology of the Cell*, vol. 20, no. 15, pp. 3503–3513, Aug. 2009, PMID: 19494041. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19494041

[55] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nature Reviews. Genetics*, vol. 10, no. 10, pp. 669–680, Oct. 2009, PMID: 19736561. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19736561

[56] Y. Gilad and J. Borevitz, "Using DNA microarrays to study natural variation," *Current Opinion in Genetics & Development*, vol. 16, no. 6, pp. 553–558, Dec. 2006, PMID: 17008090. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17008090

[57] S. Tarazona, F. Garca-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, "Differential expression in RNA-seq: a matter of depth," *Genome Research*, vol. 21, no. 12, pp. 2213–2223, Dec. 2011, PMID: 21903743. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21903743

[58] D. Lee, R. Karchin, and M. A. Beer, "Discriminative prediction of mammalian enhancers from DNA sequence," *Genome Research*, vol. 21,

no. 12, pp. 2167–2180, Dec. 2011, PMID: 21875935. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21875935

[59] T. Boes and M. Neuhäuser, "Normalization for affymetrix GeneChips," *Methods of Information in Medicine*, vol. 44, no. 3, pp. 414–417, 2005, PMID: 16113766. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16113766

[60] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics (Oxford, England)*, vol. 19, no. 2, pp. 185–193, Jan. 2003, PMID: 12538238. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12538238

[61] G. C. Tseng, M. K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong, "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects," *Nucleic Acids Research*, vol. 29, no. 12, pp. 2549–2557, Jun. 2001, PMID: 11410663. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11410663

[62] O. Thellin, W. Zorzi, B. Lakaye, B. De Borman, B. Coumans, G. Hennen, T. Grisar, A. Igout, and E. Heinen, "Housekeeping genes as internal standards: use and limits," *Journal of Biotechnology*, vol. 75, no. 2-3, pp. 291–295, Oct. 1999, PMID: 10617337. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10617337

[63] J. C. Mar, Y. Kimura, K. Schroder, K. M. Irvine, Y. Hayashizaki, H. Suzuki, D. Hume, and J. Quackenbush, "Data-driven normalization strategies for high-throughput quantitative RT-PCR," *BMC Bioinformatics*, vol. 10, p. 110, 2009, PMID: 19374774. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19374774

[64] W. Sun, M. J. Buck, M. Patel, and I. J. Davis, "Improved ChIP-chip analysis by a mixture model approach," *BMC Bioinformatics*, vol. 10, p. 173, 2009, PMID: 19500407. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19500407

[65] L. Zhang, M. F. Miles, and K. D. Aldape, "A model of molecular interactions on short oligonucleotide microarrays," *Nature Biotechnology*, vol. 21, no. 7, pp. 818–821, Jul. 2003, PMID: 12794640. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12794640

[66] W. E. Johnson, W. Li, C. A. Meyer, R. Gottardo, J. S. Carroll, M. Brown, and X. S. Liu, "Model-based analysis of tiling-arrays for ChIP-chip," *Proceedings of the National Academy of Sciences*, vol. 103, no. 33, pp. 12 457 –12 462, 2006. [Online]. Available: http://www.pnas.org/content/103/33/12457.abstract

[67] J. T. Judy and H. Ji, "TileProbe: modeling tiling array probe effects using publicly available data," *Bioinformatics (Oxford, England)*, vol. 25,

no. 18, pp. 2369–2375, Sep. 2009, PMID: 19592393. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19592393

[68] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, "NCBI GEO: mining tens of millions of expression profiles–database and tools update," *Nucleic Acids Research*, vol. 35, no. Database issue, pp. D760–765, Jan. 2007, PMID: 17099226. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17099226

[69] E. Hubbell, W. Liu, and R. Mei, "Robust estimators for expression analysis," *Bioinformatics (Oxford, England)*, vol. 18, no. 12, pp. 1585–1592, Dec. 2002, PMID: 12490442. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12490442

[70] D. S. Johnson, W. Li, D. B. Gordon, A. Bhattacharjee, B. Curry, J. Ghosh, L. Brizuela, J. S. Carroll, M. Brown, P. Flicek, C. M. Koch, I. Dunham, M. Bieda, X. Xu, P. J. Farnham, P. Kapranov, D. A. Nix, T. R. Gingeras, X. Zhang, H. Holster, N. Jiang, R. D. Green, J. S. Song, S. A. McCuine, E. Anton, L. Nguyen, N. D. Trinklein, Z. Ye, K. Ching, D. Hawkins, B. Ren, P. C. Scacheri, J. Rozowsky, A. Karpikov, G. Euskirchen, S. Weissman, M. Gerstein, M. Snyder, A. Yang, Z. Moqtaderi, H. Hirsch, H. P. Shulha, Y. Fu, Z. Weng, K. Struhl, R. M. Myers, J. D. Lieb, and X. S. Liu, "Systematic

evaluation of variability in ChIP-chip experiments using predefined DNA targets," *Genome Research*, vol. 18, no. 3, pp. 393–403, Mar. 2008. [Online]. Available: http://genome.cshlp.org/content/18/3/393

[71] A. Droit, C. Cheung, and R. Gottardo, "rMAT - an R/Bioconductor package for analyzing ChIP-Chip experiments," *Bioinformatics*, vol. 26, no. 5, pp. 678–679, Mar. 2010. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/26/5/678

[72] G. M. Santangelo, "Glucose signaling in saccharomyces cerevisiae," *Microbiol. Mol. Biol. Rev.*, vol. 70, no. 1, pp. 253–282, Mar. 2006. [Online]. Available: http://mmbr.asm.org/cgi/content/abstract/70/1/253

[73] G. Yuan, Y. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando, "Genome-scale identification of nucleosome positions in s. cerevisiae," *Science (New York, N.Y.)*, vol. 309, no. 5734, pp. 626–630, Jul. 2005, PMID: 15961632. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15961632

[74] M. Ghanbari, *Standard Codecs: Image Compression to Advanced Video Coding*. The Institution of Engineering and Technology, Jun. 2003.

[75] M. Ghandi, M. M. Ghandi, and M. B. Shamsollahi, "A novel context modeling scheme for motion vectors context-based arithmetic coding," in *Canadian*

Conference on Electrical and Computer Engineering, 2004, vol. 4. IEEE, May 2004, pp. 2021– 2024 Vol.4.

[76] N. B. Becker and R. Everaers, "DNA nanomechanics in the nucleosome," *Structure (London, England: 1993)*, vol. 17, no. 4, pp. 579–589, Apr. 2009, PMID: 19368891. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19368891

[77] D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K. Gifford, and R. A. Young, "Genome-wide map of nucleosome acetylation and methylation in yeast," *Cell*, vol. 122, no. 4, pp. 517–527, Aug. 2005, PMID: 16122420. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16122420

[78] R. E. Dickerson, "DNA bending: The prevalence of kinkiness and the virtues of normality," *Nucleic Acids Research*, vol. 26, no. 8, pp. 1906–1926, Apr. 1998. [Online]. Available: http://nar.oxfordjournals.org/content/26/8/1906

[79] W. K. Olson, A. A. Gorin, X. Lu, L. M. Hock, and V. B. Zhurkin, "DNA Sequence-Dependent deformability deduced from proteinDNA crystal complexes," *Proceedings of the National Academy of Sciences*, vol. 95, no. 19, pp. 11 163–11 168, Sep. 1998. [Online]. Available: http://www.pnas.org/content/95/19/11163

[80] A. Scipioni, C. Anselmi, G. Zuccheri, B. Samori, and P. D. Santis, "Sequence-Dependent DNA curvature and flexibility from scanning force microscopy images," *Biophysical Journal*, vol. 83, no. 5, pp. 2408–2418, Nov. 2002. [Online]. Available: http://www.cell.com/biophysj/abstract/S0006-3495(02)75254-5

[81] M. Y. Tolstorukov, A. V. Colasanti, D. M. McCandlish, W. K. Olson, and V. B. Zhurkin, "A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning," *Journal of Molecular Biology*, vol. 371, no. 3, pp. 725–738, Aug. 2007, PMID: 17585938. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17585938

[82] M. Ghorbani and F. Mohammad-Rafiee, "Geometrical correlations in the nucleosomal DNA conformation and the role of the covalent bonds rigidity," *Nucleic Acids Research*, vol. 39, no. 4, pp. 1220–1230, Mar. 2011. [Online]. Available: http://nar.oxfordjournals.org/content/39/4/1220

[83] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, "Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 resolution," *Journal of Molecular Biology*, vol. 319, no. 5, pp. 1097–1113, Jun. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283602003868

[84] X. Lu and W. K. Olson, "3DNA: a software package for the analysis, rebuilding and visualization of threedimensional nucleic acid structures," *Nucleic Acids*

Research, vol. 31, no. 17, pp. 5108–5121, Sep. 2003. [Online]. Available: http://nar.oxfordjournals.org/content/31/17/5108

[85] ——, "3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures," *Nature Protocols*, vol. 3, no. 7, pp. 1213–1227, Jul. 2008. [Online]. Available: http://www.nature.com/nprot/journal/v3/n7/abs/nprot.2008.104.html

[86] C. R. Calladine, *Understanding DNA: The Molecule & How It Works.* Academic Press, Mar. 2004.

[87] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, third edition ed. The MIT Press, Jul. 2009.

[88] G. D. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics (Oxford, England)*, vol. 16, no. 1, pp. 16–23, Jan. 2000, PMID: 10812473. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10812473

[89] P. V. Benos, M. L. Bulyk, and G. D. Stormo, "Additivity in protein-DNA interactions: how good an approximation is it?" *Nucleic Acids Research*, vol. 30, no. 20, pp. 4442–4451, Oct. 2002, PMID: 12384591. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12384591

[90] J. Göke, M. H. Schulz, J. Lasserre, and M. Vingron, "Estimation of pairwise

sequence similarity of mammalian enhancers with word neighbourhood counts," *Bioinformatics (Oxford, England)*, Jan. 2012, PMID: 22247280. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22247280

[91] J. van Helden, "Metrics for comparing regulatory sequences on the basis of pattern counts," *Bioinformatics (Oxford, England)*, vol. 20, no. 3, pp. 399–406, Feb. 2004, PMID: 14764560. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14764560

[92] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch string kernels for discriminative protein classification," *Bioinformatics (Oxford, England)*, vol. 20, no. 4, pp. 467–476, Mar. 2004, PMID: 14990442. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14990442

[93] M. R. Kantorovitz, M. Kazemian, S. Kinston, D. Miranda-Saavedra, Q. Zhu, G. E. Robinson, B. Göttgens, M. S. Halfon, and S. Sinha, "Motif-blind, genome-wide discovery of cis-regulatory modules in drosophila and mouse," *Developmental Cell*, vol. 17, no. 4, pp. 568–579, Oct. 2009, PMID: 19853570. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19853570

[94] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis, "Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals," *Nature*, vol.

434, no. 7031, pp. 338–345, Mar. 2005, PMID: 15735639. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15735639

[95] O. Elemento and S. Tavazoie, "Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach," *Genome Biology*, vol. 6, no. 2, p. R18, 2005, PMID: 15693947. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15693947

[96] P. Meinicke, M. Tech, B. Morgenstern, and R. Merkl, "Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites," *BMC Bioinformatics*, vol. 5, p. 169, Oct. 2004, PMID: 15511290. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15511290

[97] S. Sonnenburg, A. Zien, and G. Rätsch, "ARTS: accurate recognition of transcription starts in human," *Bioinformatics (Oxford, England)*, vol. 22, no. 14, pp. e472–480, Jul. 2006, PMID: 16873509. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16873509

[98] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support vector machines and kernels for computational biology," *PLoS Computational Biology*, vol. 4, no. 10, p. e1000173, Oct. 2008, PMID: 18974822. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18974822

[99] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch, "Accurate

splice site prediction using support vector machines," *BMC Bioinformatics*, vol. 8 Suppl 10, p. S7, 2007, PMID: 18269701. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18269701

[100] A. E. Albert, *Regression and the Moore-Penrose pseudoinverse.* Academic Press, Oct. 1972.

[101] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, 2nd ed. Addison-Wesley Professional, Mar. 1994.

[102] R. M. Wilson, "A diagonal form for the incidence matrices of <italic>t</italic>-subsets vs. <italic>k</italic>-subsets," *Eur. J. Comb.*, vol. 11, no. 6, p. 609615, Oct. 1990. [Online]. Available: http://dl.acm.org/citation.cfm?id=107902.107915

[103] M. Ghandi, M. Mohammad-Noori, and M. A. Beer, "Robust k-mer frequency estimation using gapped k-mers," *Jourbal of Mathematical Biology – under review*, 2012.

[104] R. Sandberg, G. Winberg, C. Bränden, A. Kaske, I. Ernberg, and J. Cöster, "Capturing Whole-Genome characteristics in short sequences using a nave bayesian classifier," *Genome Research*, vol. 11, no. 8, pp. 1404–1409, Aug. 2001. [Online]. Available: http://genome.cshlp.org/content/11/8/1404

[105] W. Fury, F. Batliwalla, P. K. Gregersen, and W. Li, "Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion," *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, vol. 1, pp. 5531–5534, 2006, PMID: 17947148. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17947148

[106] R. McDaniell, B. Lee, L. Song, Z. Liu, A. P. Boyle, M. R. Erdos, L. J. Scott, M. A. Morken, K. S. Kucera, A. Battenhouse, D. Keefe, F. S. Collins, H. F. Willard, J. D. Lieb, T. S. Furey, G. E. Crawford, V. R. Iyer, and E. Birney, "Heritable Individual-Specific and Allele-Specific chromatin signatures in humans," *Science*, vol. 328, no. 5975, pp. 235–239, Apr. 2010. [Online]. Available: http://www.sciencemag.org/content/328/5975/235.short

[107] A. Visel, M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin, and L. A. Pennacchio, "ChIP-seq accurately predicts tissue-specific activity of enhancers," *Nature*, vol. 457, no. 7231, pp. 854–858, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1038/nature07730

[108] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S.

Liu, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biology*, vol. 9, no. 9, p. R137, 2008, PMID: 18798982. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18798982

[109] T. Jaakkola, M. Diekhans, and D. Haussler, "A discriminative framework for detecting remote protein homologies," *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, vol. 7, no. 1-2, pp. 95–114, Apr. 2000, PMID: 10890390. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10890390

[110] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in r," *Bioinformatics (Oxford, England)*, vol. 21, no. 20, pp. 3940–3941, Oct. 2005, PMID: 16096348. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16096348

[111] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 2, pp. 28–36, 1994, PMID: 7584402. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7584402

[112] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed. SIAM: Society for Industrial and Applied Mathematics, Aug. 2002.

[113] N. Kaplan, I. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, "Nucleosome sequence preferences influence in vivo nucleosome organization," *Nat Struct Mol Biol*, vol. 17, no. 8, pp. 918–920, 2010. [Online]. Available: http://dx.doi.org/10.1038/nsmb0810-918

[114] G. Badis, E. T. Chan, H. van Bakel, L. Pena-Castillo, D. Tillo, K. Tsui, C. D. Carlson, A. J. Gossett, M. J. Hasinoff, C. L. Warren, M. Gebbia, S. Talukder, A. Yang, S. Mnaimneh, D. Terterov, D. Coburn, A. Li Yeo, Z. X. Yeo, N. D. Clarke, J. D. Lieb, A. Z. Ansari, C. Nislow, and T. R. Hughes, "A library of yeast transcription factor motifs reveals a widespread function for rsc3 in targeting nucleosome exclusion at promoters," *Molecular Cell*, vol. 32, no. 6, pp. 878–887, Dec. 2008, PMID: 19111667. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19111667

[115] M. Ganapathi, M. J. Palumbo, S. A. Ansari, Q. He, K. Tsui, C. Nislow, and R. H. Morse, "Extensive role of the general regulatory factors, abf1 and rap1, in determining genome-wide chromatin structure in budding yeast," *Nucleic Acids Research*, Nov. 2010, PMID: 21081559. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21081559

[116] J. T. Arnone and M. A. McAlear, "Adjacent gene pairing plays a role in the coordinated expression of ribosome biogenesis genes MPP10 and YJR003C in

saccharomyces cerevisiae," *Eukaryotic Cell*, vol. 10, no. 1, pp. 43–53, Jan. 2011. [Online]. Available: http://ec.asm.org/cgi/content/abstract/10/1/43

[117] S. Ozcan and M. Johnston, "Three different regulatory mechanisms enable yeast hexose transporter (HXT) genes to be induced by different levels of glucose," *Molecular and Cellular Biology*, vol. 15, no. 3, pp. 1564–1572, Mar. 1995, PMID: 7862149. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7862149

[118] M. J. Brauer, C. Huttenhower, E. M. Airoldi, R. Rosenstein, J. C. Matese, D. Gresham, V. M. Boer, O. G. Troyanskaya, and D. Botstein, "Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast," *Molecular Biology of the Cell*, vol. 19, no. 1, pp. 352–367, Jan. 2008, PMID: 17959824 PMCID: 2174172.

[119] S. Busti, P. Coccetti, L. Alberghina, and M. Vanoni, "Glucose Signaling-Mediated coordination of cell growth and cell cycle in saccharomyces cerevisiae," *Sensors*, vol. 10, no. 6, pp. 6195–6240, Jun. 2010. [Online]. Available: http://www.mdpi.com/1424-8220/10/6/6195/

[120] S. Ozcan and M. Johnston, "Function and regulation of yeast hexose transporters," *Microbiology and Molecular Biology Reviews: MMBR*, vol. 63, no. 3, pp. 554–569, Sep. 1999, PMID: 10477308. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10477308

[121] J. Kim, V. Brachet, H. Moriya, and M. Johnston, "Integration of

transcriptional and posttranslational regulation in a glucose signal transduction pathway in saccharomyces cerevisiae," *Eukaryotic Cell*, vol. 5, no. 1, pp. 167–173, Jan. 2006, PMID: 16400179. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16400179

[122] J. Kim, "DNA-binding properties of the yeast rgt1 repressor," *Biochimie*, vol. 91, no. 2, pp. 300–303, Feb. 2009, PMID: 18950675 PMCID: 2859070.

[123] J. Kim and M. Johnston, "Two glucose-sensing pathways converge on rgt1 to regulate expression of glucose transporter genes in saccharomyces cerevisiae," *J. Biol. Chem.*, vol. 281, no. 36, pp. 26 144–26 149, 2006. [Online]. Available: http://dx.doi.org/10.1074/jbc.M603636200

[124] J. Kim, J. Polish, and M. Johnston, "Specificity and regulation of DNA binding by the yeast glucose transporter gene repressor rgt1," *Molecular and Cellular Biology*, vol. 23, no. 15, pp. 5208–5216, Aug. 2003, PMID: 12861007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12861007

[125] J. A. Polish, J. Kim, and M. Johnston, "How the rgt1 transcription factor of saccharomyces cerevisiae is regulated by glucose," *Genetics*, vol. 169, no. 2, pp. 583–594, Feb. 2005, PMID: 15489524. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15489524

[126] M. Floer, X. Wang, V. Prabhu, G. Berrozpe, S. Narayan, D. Spagna, D. Alvarez, J. Kendall, A. Krasnitz, A. Stepansky, J. Hicks, G. O.

Bryant, and M. Ptashne, "A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding," *Cell*, vol. 141, no. 3, pp. 407–418, Apr. 2010, PMID: 20434983. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20434983

[127] B. P. Cormack, R. H. Valdivia, and S. Falkow, "FACS-optimized mutants of the green fluorescent protein (GFP)," *Gene*, vol. 173, no. 1 Spec No, pp. 33–38, 1996, PMID: 8707053. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8707053

[128] O. Hobert, "PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic c. elegans," *BioTechniques*, vol. 32, no. 4, pp. 728–730, Apr. 2002, PMID: 11962590. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11962590

[129] M. M. Ling and B. H. Robinson, "Approaches to DNA mutagenesis: An overview," *Analytical Biochemistry*, vol. 254, no. 2, pp. 157–178, Dec. 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003269797924283

[130] W. A. Kibbe, "OligoCalc: an online oligonucleotide properties calculator," *Nucleic Acids Research*, vol. 35, no. Web Server, pp. W43–W46, May 2007. [Online]. Available: http://nar.oxfordjournals.org/content/35/suppl_2/W43.abstract

[131] R. D. Gietz, R. H. Schiestl, A. R. Willems, and R. A. Woods, "Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure," *Yeast (Chichester, England)*, vol. 11, no. 4, pp. 355–360, Apr. 1995, PMID: 7785336. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7785336

[132] J. D. Boeke, J. Trueheart, G. Natsoulis, and G. R. Fink, "5-Fluoroorotic acid as a selective agent in yeast molecular genetics," *Methods in enzymology*, vol. 154, pp. 164–175, 1987, PMID: 3323810. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3323810

[133] J. M. Chambers and T. J. Hastie, Eds., *Statistical Models in S.* Chapman and Hall/CRC, Oct. 1991.

[134] C. E. Paquin and V. M. Williamson, "Ty insertions at two loci account for most of the spontaneous antimycin a resistance mutations during growth at 15 degrees c of saccharomyces cerevisiae strains lacking ADH1," *Molecular and cellular biology*, vol. 6, no. 1, pp. 70–79, Jan. 1986, PMID: 3023838. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3023838

[135] A. Wach, A. Brachat, C. Alberti-Segui, C. Rebischung, and P. Philippsen, "Heterologous HIS3 marker and GFP reporter modules for PCR-targeting in saccharomyces cerevisiae," *Yeast (Chichester, England)*, vol. 13, no. 11, pp. 1065–1075, Sep. 1997, PMID: 9290211. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9290211

[136] H. H. Ng, F. Robert, R. A. Young, and K. Struhl, "Genome-wide location and regulated recruitment of the RSC nucleosome-remodeling complex," *Genes & Development*, vol. 16, no. 7, pp. 806–819, Apr. 2002, PMID: 11937489. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11937489

VITA

MAHMOUD GHANDI

Education:

Johns Hopkins University School of Medicine, Baltimore, MD
Ph.D, Biomedical Engineering 2006-2012

Sharif University of Technology, Tehran, Iran
M. Sc., Electrical Engineering 2003-2006

Sharif University of Technology, Tehran, Iran
B. Sc., Electrical Engineering 1999-2003

Awards:

Programming Excellence Award, Upsilon Pi Epsilon 2001

Silver medal ACM ICPC world finals 2001

Silver Medal, 11th International Olympiad in Informatics 1999

Peer-Reviewed Publications:

M. Ghandi, M. M-Noori and M. Beer, 'Robust k-mer Frequency Estimation Using Gapped k-mers,' submitted to Journal of Mathematical Biology, Jan 2012.

M. Ghandi and M. Beer, 'Group Normalization for Genomic Data,' accepted for publication in PLoS ONE, Jun 2012.

Y. Jing, Y. Qian, **M. Ghandi**, A. He, S. Pan, Z. Jian, 'A Mechanistic Study on the Effect of Dexamethasone in Moderating Cell Death in Chinese Hamster Ovary Cell Cultures,' Biotechnol. Prog, Nov 2011.

M. Ghandi, M. Yekta, F. Marvasti, 'Some Nonlinear/Adaptive Methods for Fast Recovery of the Missing Samples of Signals,' Elsevier Journal of Signal Processing, vol. 88, issue 3, Mar 2008.