Aalto University
School of Science
Degree Programme in Computer Science and Engineering

Basak Eraslan

A probabilistic model for competitive DNA binding modeling using ChIP-seq and MNase-seq data

Master's Thesis
Espoo, April 23, 2015

Supervisors: Professor Harri Lähdesmäki
Professor Erik Aurell
Advisor: Professor Harri Lähdesmäki
Henrik Mannerström

Aalto University
School of Science
Degree Programme in Computer Science and Engineering

ABSTRACT OF
MASTER'S THESIS

Author: Basak Eraslan

Title:
A probabilistic model for competitive DNA binding modeling using ChIP-seq and MNase-seq data

Date:	April 23, 2015	**Pages:**	48
Major:	Computational Systems Biology	**Code:**	T-61
Supervisors:	Professor Harri Lähdesmäki		
Advisor:	Professor Harri Lähdesmäki Henrik Mannerström		

Competitive and combinatorial DNA binding pattern of transcription factors and nucleosomes at genomic regulatory regions control the key cellular processes such as transcription, replication and chromatin packaging. Consequently, in order to reveal the gene expression regulatory mechanisms, it is critical that we understand how these DNA binding factors (DBFs) are organized in the cell under specific conditions. The quantitative models proposed for predicting the complex combinatorial binding pattern underlying gene expression generally use the DNA binding affinities and concentrations of the DNA binding factors. These models have been shown to work well under thermodynamic equilibrium conditions in lower organisms but when modeling the actual in vivo binding we have to consider the ATP-driven chromatin remodelers actively repositioning, reconfiguring or ejecting nucleosomes, the binding cooperativity among transcription factors and the environment of the cell with ATP-driven molecular components acting against thermal equilibrium. Moreover, the challenge of correctly determining DBF concentrations in the cell makes the application of these methods troublesome. In this study, we propose a probabilistic method to infer the competitive and combinatorial DNA occupancy of the factors at each position of an inspected region by the use of the ChIP-Seq and MNase-Seq high-throughput data which intrinsically reflect the effects of all of the factors related with DBF positioning. Our method is built upon the enriched read coverage profiles observed around the binding sites and explicitly includes the competition between DBFs. Experiments we have conducted with 47 DBFs suggest that incorporation of this competition into the model increases the precision of the binding site estimates.

Keywords: transcription factors, nucleosomes, competitive binding, Chip-seq, MNase-seq, ENCODE

Language: English

Acknowledgements

I would like to thank Prof. Harri Lähdesmäki and Mr. Henrik Mannerström for their great support and guidance throughout this study.

Espoo, April 23, 2015

Basak Eraslan

Abbreviations and Acronyms

ChIP-Chip	Chromatin immunoprecipitation with DNA microarray
ChIP-Seq	Chromatin immunoprecipitation with massively parallel DNA sequencing
DBF	DNA binding factor
ENCODE	The encyclopedia of DNA elements
MNase	Micrococcal nuclease
TF	Transcription factor
TSS	Transcription start site

Contents

Chapter 1

Introduction

Regulation of the key eukaryotic cellular processes like transcription, replication and chromatin packaging is dependent on the binding of hundreds of different factors to the genome. These proteins include histone proteins and transcription factors.

Eukaryotic genome needs to fit in the small volume of the nucleus while being accessible to the DNA binding factors that control gene expression. Nucleosomes, which occupy $\sim 75-90$ of the genome [9], are the basic building blocks of the chromatin structure which is specially evolved to achieve in this goal. Nucleosomes are formed by approximately 147bps of DNA wrapping around a histone octamer, which contains two copies of each of the core histones: H2A, H2B, H3, and H4 [10]. The 10-50bp long stretches of DNA that run between nucleosomes are referred to as linker DNA. Figure 1.1 shows the hierarchical chromatin structure. The linear arrangement of the nucleosomes along the DNA constitutes the primary packing level. With addition of H1 histone proteins, multiple histones wrap into helical structures called chromatin fibers which eventually form chromosomes [7].

Transcription factors (TFs) are DNA binding proteins which promote or inhibit gene expression. The net expression outcome is dependent on the concentrations and binding affinities of these factors in the cell [8]. DNA binding domains of the TFs are specialized in recognizing and binding energetically preferable locations of the genome which are called as binding sites. The proteins which lack DNA binding domains, e.g. coactivators, chromatin remodelers, histone acetylases, deacetylases, kinases, methylases, are not labelled as transcription factors even though they might play important roles in gene expression regulation [11].

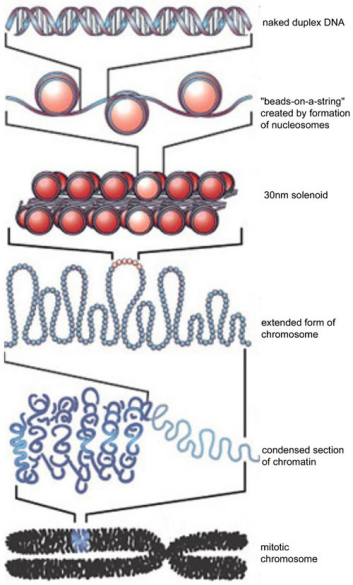


Figure 1.1: Chromatin structure which enables eukaryotic genomic DNA to fit into the nucleus of the cell while being accessible to DNA binding factors that control gene expression. DNA is wrapped around histone octamers form nucleosomes which are connected by linker DNA [35].

In general, the compactness of the chromatin is inversely proportional to DNA accessibility of the DNA binding factors. In other words, the more tightly packaged the DNA, the harder it is for the transcription factors and other DNA binding proteins to access DNA and perform their functions [7]. Even though some examples have been reported for configurations where TFs are able to bind to nucleosomal DNA [1], in most cases TFs cannot bind nucleosomal sequences and hence compete with nucleosomes for DNA access. Additionally, binding sites of several TFs that control the expression of the same gene may be overlapping. This again motivates a competition between different TFs. Consequently, in addition to the TF binding preferences, the sites to which TFs can bind also depend on the competitions between TFs and nucleosomes as well as the competitions among TFs[2]. Cellular concentrations and binding affinities of these factors generally determine the winners of these races. Likewise, nucleosome organization is determined by multiple factors, including the competition with these site specific DNA-binding proteins, DNA sequence preferences of the nucleosomes and active chromatin remodellers that reposition or remove nucleosomes [12, 72]. Due to the competitive binding between TFs and nucleosomes, the organization of nucleosomes play a significant role on transcriptional gene expression regulation.

1.1 TF Binding Discovery and Prediction

Many studies have been conducted on both transcription factor binding site (TFBS) discovery and prediction, and nucleosome positioning. Most of the computational approaches to TFBS analysis employ position weight matrices (PWMs) which quantitatively represent TF binding motifs [13](see [14] for a review). However, since TFBSs are usually short and the sequence changes at many positions of the binding sites are generally tolerated by TFs, with these methods suitable sequences are found excessively leading to low accuracy predictions. It is even claimed that, in a human cell most computationally predicted TFBS are not available for binding [13]. Many methods that point out this high false positive rate have been proposed in the literature [15–19] and more flexible models are suggested for the prediction of TFBSs [20–25].

Other than the computational methods, high-throughput experimental methods such as chromatin immunoprecipitation with DNA microarray (ChIP-Chip) [26, 27], chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-Seq) [28] have been developed for mapping of protein binding events. Since an array is restrained to a fixed number of probes, ChIP-Seq has gained popularity over ChIP-Chip as the sequencing cost has decreased over time.Currently ChIP-seq is the major TF binding mapping method used in the ENCODE project. Figure 1.2 displays the ChIP-Seq workflow.

In the ChIP-Seq protocol, first the target protein is cross-linked with the DNA site it binds to. Then, the cells are lysed and the DNA is sheared by sonication or by using endonuclease enzymes. This brings about chunks of protein-DNA complexes, where the double-stranded DNA is generally 1 kb or less in length [31]. In the next step, only the complexes with the protein of interest are filtered out by using an antibody specific to the target protein. The cross-linking of protein-DNA complexes is reversed and the DNA strands are purified. Oligonucleotide adaptors are then added to the small stretches of DNA that were bound to the target protein to enable massively parallel sequencing. Finally, after size selection, all the resulting ChIP-DNA fragments are sequenced simultaneously using a genome sequencer [31].

In principle, a pool of DNA fragments enriched for the target protein's binding sites should be produced at the end of the ChIP protocol. High-throughput sequencing of these fragments generates millions of short tags which are later mapped to the reference genome [33]. Due to the repetitive regions of the genome, some of these tags may be mapped to multiple locations. This should be taken into account during the downstream analysis.

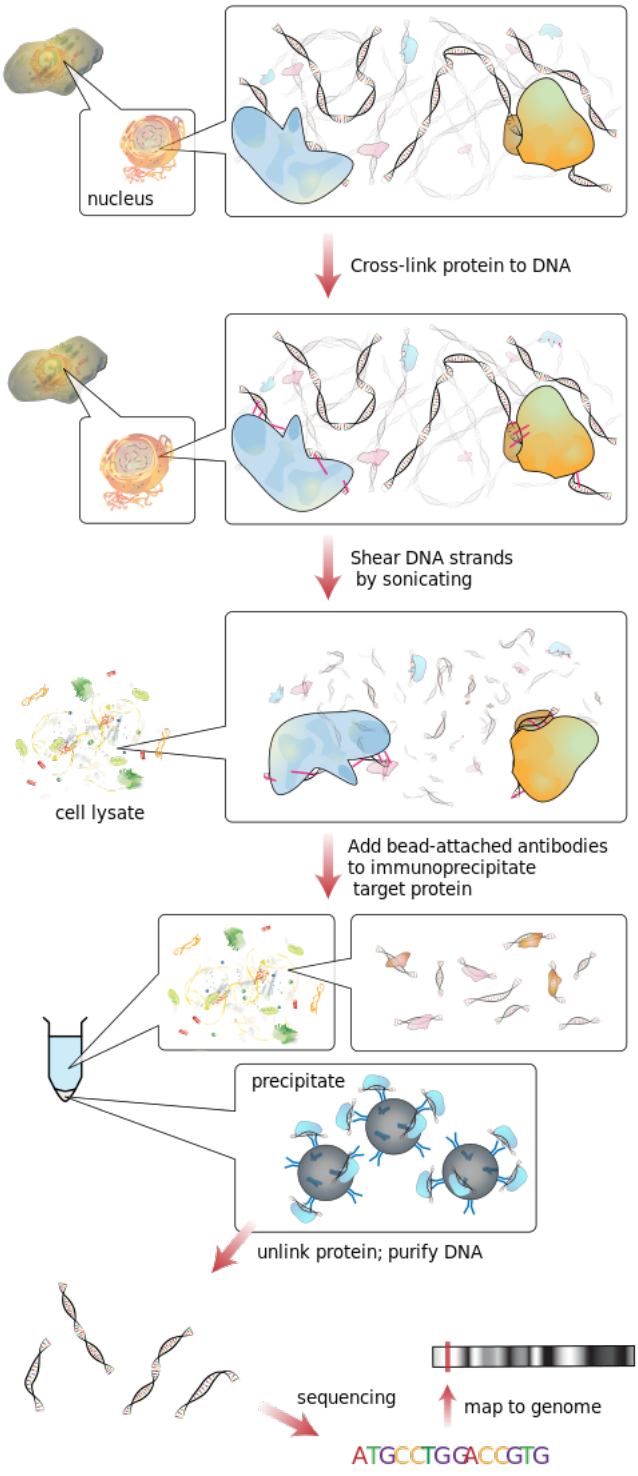


Figure 1.2: ChIP-Seq workflow [30].

For single-end sequencing, the average fragment size is estimated by the distribution of distances between reads on the positive and negative strands and each read is extended to the average fragment size, while for paired-end sequencing, the distance between the paired-end reads defines the actual length of each fragment. In order to identify the protein binding sites, ChIP-Sek peak finders identify genomic locations where mapped sequence tags are enriched and specify some criteria to distinguish the significantly enriched sites. Various approaches has been proposed for this task, e.g., identifying regions where extended sequence tags overlap, sliding windows algorithm for finding fixed width windows in which the number of tags are enriched, and searching bimodal pattern in the strand-specific tag densities (for a review of peak calling programs see [33]).

1.2 Nucleosome Positioning

Non overlapping positions of the nucleosomes on the DNA constitutes the nucleosome configuration of a cell[36]. Higher organisms have varying nucleosome configurations in different cell types and even in a specific cell sample, the exact positions of the nucleosomes within each cell may deviate around a most preferred position. This variance of nucleosome positions within each cell in a cell population is referred to as fuzziness [37]. Nucleosome positioning aims to assess the most preferred positions, fuzziness and the occupancy values of the individual nucleosomes. Occupancy value of a nucleosome indicates the frequency with which this genome position is occupied by a nucleosome in a cell population [37]. If a nucleosome is "well-positioned", this means that the nucleosome is present at the same genomic location in most of the cells in the population.

There are many factors determining the nucleosome positions. These include the sequence characteristics of the local DNA and the intrinsic sequence preferences of the nucleosomes, active chromatin complexes that reposition or delete nucleosomes, competition with other DNA binding factors and the effects of neighboring nucleosomes (for a review of cis and trans determinants of nucleosome positioning see [38]). In a cell nucleosome positions can be altered as a response to dynamic environmental factors such as heat shock or hormonal treatment [37]. Therefore, accurate positioning of nucleosomes and studying the nucleosome repositioning mechanisms is a key to understand how chromatin elements and transcription factors collaborate to organize the cellular responses to environmental changes [39, 40].

In order to wrap around the histone octamer the DNA helix is has to experience a sharp bending. Certain sequences are argued to intrinsically favor or disfavor this curving. Nucleosomes are more likely to be formed when the bending is preferred [45, 46]. Based on this observation, numerous studies have predicted in vivo nucleosome positions directly from DNA sequence, claiming that nucleosome organizations are encoded in the genomic sequence to a certain extent [41–44]. With the advance of high-throughput sequencing, new experimental techniques have been designed to measure nucleosome positions on a genome-wide scale.

Micrococcal nuclease (MNase) digests chromatin at DNA sites that are not occupied by nucleosomes, because linker DNA in between nucleosomes is more exposed to the nuclease while nucleosome covered sequences are better protected from digestion [48]. Figure 1.3 illustrates the digestion process. This step introduces some biases to the experiment results, because MNase has a sequence preference to having TA/AT dinucleotide as its cleavage site [49, 50]. Fortunately, it has been demonstrated that nucleosome protection is more effective than MNase specificity in the MNase digestion [36]. After digestion, the beads are isolated and the nucleosomal DNA is extracted for high-throughput sequencing. The sequenced reads are again mapped to the reference genome, followed by some normalization steps. Nucleosomes are then positioned according to the peaks of the coverage profile, just as in the ChIP-seq peak calling process. However, the resulting coverage profile generally exhibits many "blurry" peaks which involves overlapping and ambiguous nucleosome positions [34]. This is due to three main reasons: firstly, MNase sequence preferences cause some DNA fragment length variation from the actual nucleosomal DNA size, which is hypothetically 147bp [54]. Secondly, considering that MNase is a strong enzyme, the digestion outcome is very easily affected by the MNase concentration and incubation time [51–53]. Thirdly, since it is not possible to locate the nucleosomes cell by cell with the current technologies, the deviation in the nucleosome positions between cells is directly reflected in the experimental results obtained from a population of cells[54]. Several approaches and tools have been proposed to identify nucleosome positions from MNase-Seq data, such as peak calling [55, 56], and hidden Markov models [44] (for a recent review of nucleosome positioning methods see [60]).

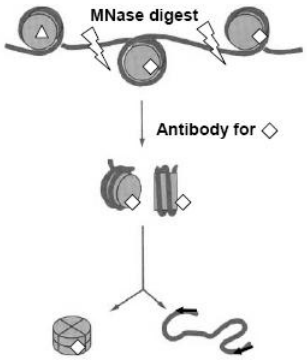


Figure 1.3: MNase digests chromatin at DNA sites that are not occupied by nucleosomes [47].

1.3 Problem statement

Hypothetically any DNA sequence of a certain length is a candidate binding site for a DNA binding factor. The binding probability of a factor to a site is a time-dependent dynamic variable determined by the sequence preferences of the factor, its concentration in the cell at that specific time and the stochastic competitions between different factors. Therefore, it would be a more realistic view of the DNA-protein interactions if the genomic positions were annotated as binding sites using a probability interval. However, current available models generally announce the positions to be either binding sites or not[2].

Secondly, most of the models of genome binding do not consider different types of factors together simultaneously even though considering the competition between different factors would yield more accurate locations for both the nucleosomes and the TF binding sites [2]. In addition, methods that model the competitive and combinatorial binding pattern of transcription factors and nucleosomes at genomic regulatory regions would be beneficial for revealing the complex mechanisms of gene expression regulation. The quantitative models proposed for predicting this complex combinatorial code generally use the binding affinities and concentrations of the DBFs in the cell [1–6]. These models might work well in lower organisms under thermody-

namic equilibrium conditions but when modeling actual in vivo binding we have to consider the ATP-driven chromatin remodelers capable of actively repositioning, reconfiguring or ejecting nucleosomes [7], the binding cooperativity among transcription factors [8] and the environment of the cell with ATP-driven molecular components functioning against thermal equilibrium [57]. Moreover, the challenge of correctly determining DBF concentrations in the cell makes the application of these methods troublesome [2].

In this study we propose a probabilistic model for generating the combinatorial "occupancy profile" [2] of TFs and nucleosomes in a cell sample. Our model uses Chip-Seq and MNase-Seq data sets of the TFs and nucleosomes as input which intrinsically reflects the direct and indirect effects of the factors related with DBF positioning. To our knowledge, ours is the first study which explicitly models the combinatorial configuration of multiple DBFs by the use of the high-throughput sequencing data.

1.4 Structure of the Thesis

Chapter 2 represents the materials and the enrichment analysis we performed on ENCODE ChIP-Seq and MNase-Seq GM12878 cell line data which is prepared by using human lymphoblastoid cell type. Chapter 3 explains our proposed method in detail and Chapter 4 displays the experiment results. Finally Chapter 5 discusses about our findings and concludes the study.

Chapter 2

Materials

In our experiments we used the publicly available GM12878 cell line human ChIP-Seq and MNase-Seq data sets released by the ENCODE project [58]. GM12878 is a lymphoblastoid cell line produced from the blood of a female donor with northern and western European ancestry by EBV transformation [59]. It is one of the two cell lines (other one is K562) for which MNase-Seq data is published. Currently there are 98 TFs for which ChIP-Seq data is published in this cell line. The 46 TFs that we used in our experiments are: ATF2, ATF3, BCL3, BCL11A, BCLAF1, CEBPB, CREB1, CTCF, EBF1, EGR1, ELF1, ETS1, FOXM1, GABPA, IRF4, MEF2A, MEF2C, MTA3, NFATC1, NFE2, NFIC, NRF1, NRSF, PAX5, PBX3, PML, POL2, POL24H8, POU2F2, PU1, RAD21, RFX5, RUNX3, RXRA, SIX5, SP1, STAT1, STAT5A, TAF1, TBP, TCF3, TCF12, USF1, YY1, ZBTB33, ZEB1.

2.1 Data Preprocessing

2.1.1 ChIP-Seq Data Preprocessing

We downloaded raw ChIP-Seq data files from UCSC downloads server [61] in .fastq format and used FastQC [62] to perform quality control checks to ensure that the raw data looks good and there are no problems or biases in our data. FastQC provides a QC report which can spot problems which originate either in the sequencer or in the starting library material. Figure 2.1 displays an example FastQC report summary for the raw data file of TF RUNX3. Most of the TFs we used had high quality data according to FastQC analysis reports. For some TFs, there were overrepresented sequences which were reported to be the adapter sequences.

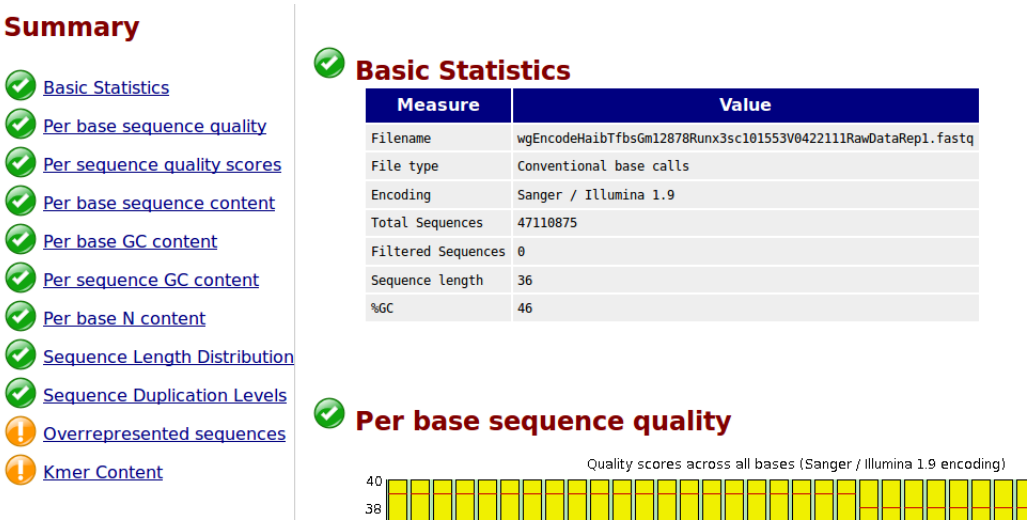


Figure 2.1: FastQC report summary example.

We removed such sequences that were ligated to the 5' or 3' ends of the library reads by the use of Cutadapt [63] software with the "-q 20 -m 36" options which enabled us to trim low-quality ends from reads before adapter removal and discard trimmed reads that are shorter than 36bps.

After quality check and adapter removal, we aligned the library reads to the reference genome NCBI GRCh37 (hg19) with the Bowtie software [64]. Reads that were mapped to more than one single location were discarded. Following that, first we converted resulting .sam format files into .bam files with "Samtools view" software [65] and then used "Samtools rmdup" to remove the duplicate reads. Duplicate removal is important for handling the possible amplification and sequencing errors [55]. Finally, we converted the duplicate removed .bam files into .bed format which is the input file format of our tool.

In our model we use the reads in 1bp resolution. In other words, each read is represented by its 5' end coordinate. Samtools only retains the read with highest mapping quality if multiple reads have identical coordinates for their 5' and 3' ends and the reads with identical 5' end positions are kept when their 3' ends are different. Considering the trimming steps in our preprocessing procedure, we only keep one of the reads which have identical 5' positions and discard the others. Therefore, the occupancy value of each base pair is at most 2 (1 read mapped to the positive strand and 1 read mapped to the negative strand).

Inputs to our implementation contain the MACS summit reports for the considered TFs. In each of these reports, in addition to the peak summit locations, MACS announces the average fragment length of the ChIP-seq

library. We read that value and shift the reads in the 3' direction by half of the fragment length to locate the center of each fragment. As a last step, the shifted 1bp resolution reads are binned (default bin size is 5bp) and the number of reads in each bin is used in the model.

2.1.2 MNase-Seq Data Preprocessing

During sonication or endonuclease digestion some regions of open chromatin are preferentially cut. This preference and errors during sequencing generate fragments which are mapped to non-specific background regions throughout the genome. Therefore, the reads in a ChIP sample contains some background noise reads as well as the enrichment signal reads. In order to reduce the effect of the noise in the data, the experiment signal should be analyzed considering a reference sample. There are two basic methods to generate the control data for the ChIP-seq experiments. Both of these methods use the cells from the same sample as used in the ChIP sample. In the first method, the "input DNA" is produced by cross linking and fragmenting the cell DNA without immunoprecipitation. The second method includes immunoprecipitation step but uses an antibody that has no specificity to any protein. The control data obtained by following the second technique is called "IgG" control. Input control ChIP-seq data is published for the ChIP-seq that we used in our experiments. We preprocess the control data with the same steps used in preprocessing the sample data. In order to normalize the control data, we calculate the ratio between the total control tag count and total sample tag count, and multiply the binned control signal with this ratio before subtracting it from the binned sample signal. -

For the preprocessing of ENCODE GM12878 cell line MNase-Seq data, we followed the same steps explained in Section 2.1.1, except that for this data set reads and index were in colorspace and necessary arguments ("-c" for cutadapt and "-C" for bowtie) were specified to align the reads to the reference genome hg19.

2.2 Data Enrichment Analysis

2.2.1 ChIP-Seq Data Enrichment Analysis

We performed ChIP-Seq read enrichment analysis around the peak summits detected by MACS [66] for all of the TFs we considered throughout our experiments. When we inspected the read distributions in 800bp regions centered at 400 most significant peak summits of MACS, we observed that the number

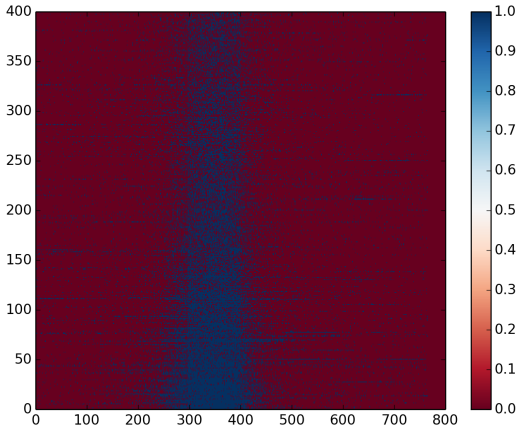
of reads mapped to the + strand are enriched in a $[-w, 0]$ window and the number of reads mapped to the - strand are enriched in a $[0, w]$ window where w was a number close to the average fragment length reported by MACS. Therefore, when we shifted the 5' end positions of the reads by $w/2$, the cumulative signal in a w bp window centered at the summits was significantly enriched compared to the background. This observation is compatible with the assumptions of most of the peak calling algorithms [33]. As an example of these analyses, Figure 2.2 displays the heat maps for the 1bp resolution read distribution in 800bp regions centered at 400 most significant EBF1 peak summits on chromosome one that were reported by MACS and Figure 2.3 shows the average density profiles of these 400 regions. MACS reports the average fragment length of this TF ChIP-Seq library to be 103bp, and accordingly we see that there is a significant enrichment in a $\sim \pm 100$ region centered at the summits.

2.2.2 MNase-Seq Data Enrichment Analysis

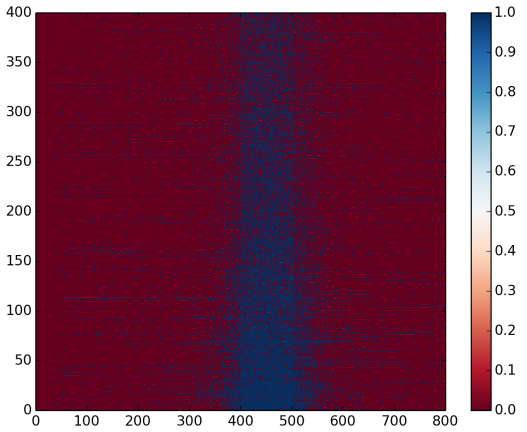
We also performed a similar analysis on the MNase-Seq data. For that purpose, we ran DANPOS [37] software on MNase-Seq data for ENCODE GM12878 cell line, chromosome one and after ordering the nucleosome summits according to their p-values, we analyzed the 1bp resolution read distribution at the regions which are centered at 500 most significant summit positions.

Although nucleosome size is 147bp in higher eukaryotes, due to the noisy nature of existing nucleosome positioning data, the real size of DNA fragments after MNase digestion vary from $\sim$ 120bp to 210bp [12, 37]. Moreover, MNase-Seq data is more scattered compared with Chip-Seq data because of the mobility of the nucleosomes in a cell population. Since we had single-end sequenced MNase-Seq data, in order to find the average fragment length, we measured the phase shift between the + strand read density and - strand read density by calculating the cross correlation between + and - density signals.

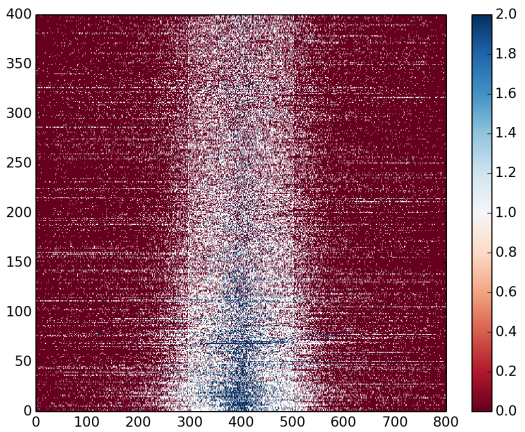
Figure 2.4 displays the heat maps for 1bp resolution read distribution in 600bp regions centered at 500 significant nucleosome summit positions detected by DANPOS and Figure 2.5 shows average 1bp read density profiles of these regions. In Figure 2.5a, we observe the phase-shift value between + and - strand density signals to be $\sim$ 200bp. When the reads were shifted by 100bp, the enriched regions of the two signals overlap as shown in Figure 2.5b. In this figure, we see that the number of fragment centers of MNAse-Seq reads are enriched in a $\sim$ 75bp window. These MNase-Seq enrichment analysis results are compatible with the findings published in [55].



(a) Reads mapped to + strand.

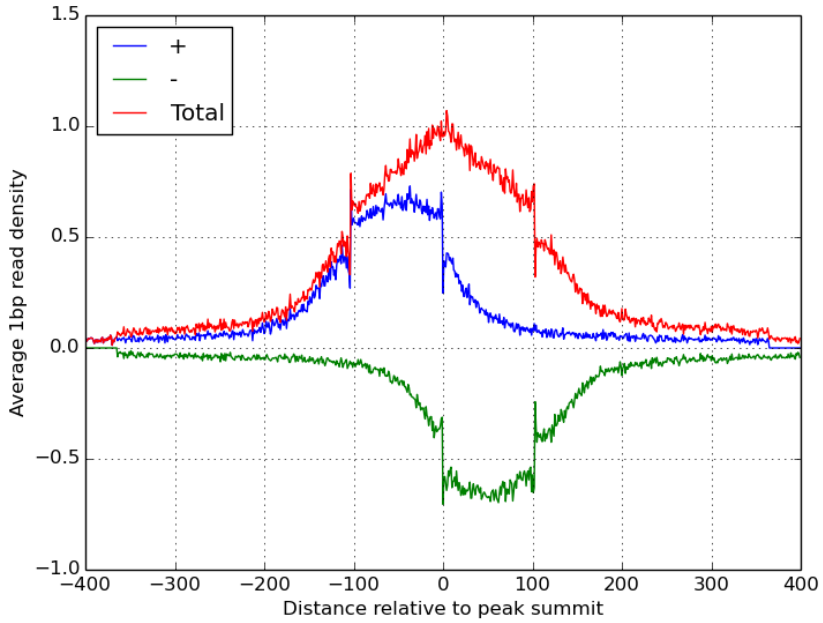


(b) Reads mapped to - strand.

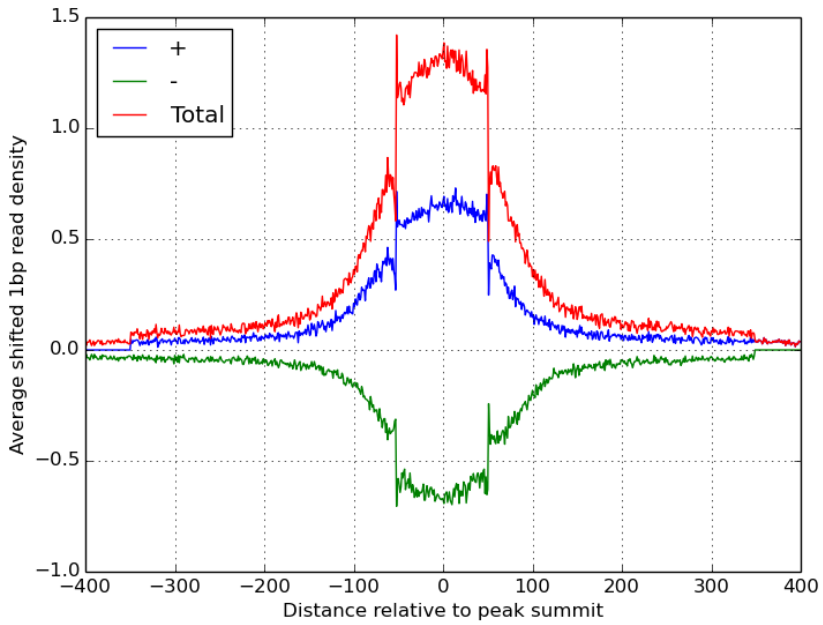


(c) Reads mapped to both strands.

Figure 2.2: Heat maps displaying the 1bp resolution read distribution in 800bp regions centered at peak summits (most significant 400) for EBF1 protein detected by MACS.

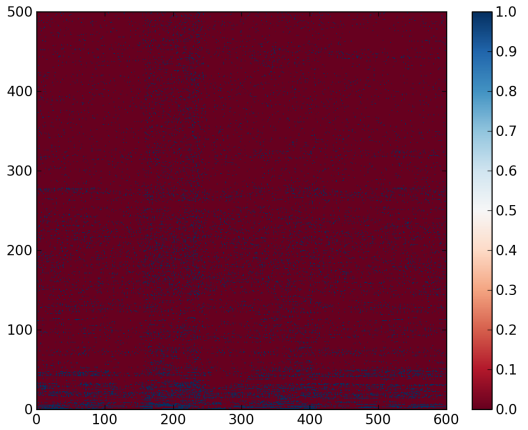


(a) Fragments are unshifted.

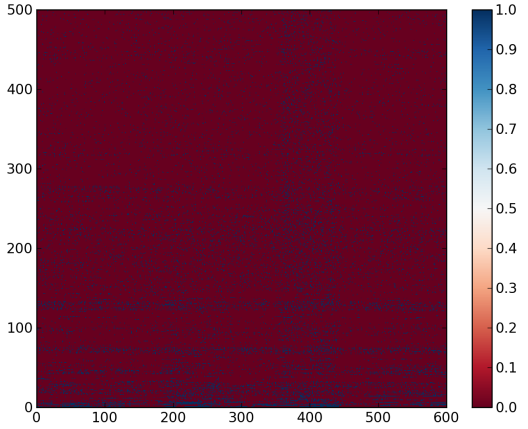


(b) Fragments are shifted by $\frac{d}{2}$, which is 51bp for EBF1.

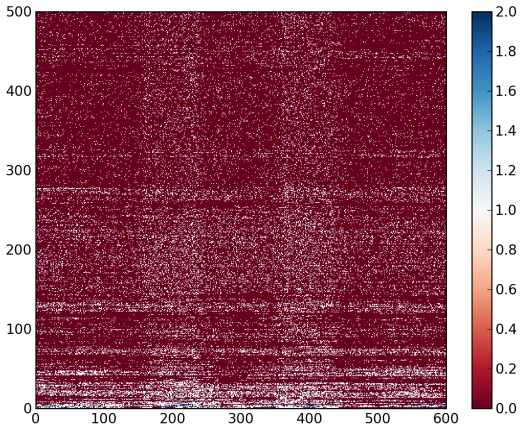
Figure 2.3: Average 1bp resolution read density profiles in 800bp regions centered at MACS reported peaks summits for EBF1.



(a) Reads mapped to + strand.

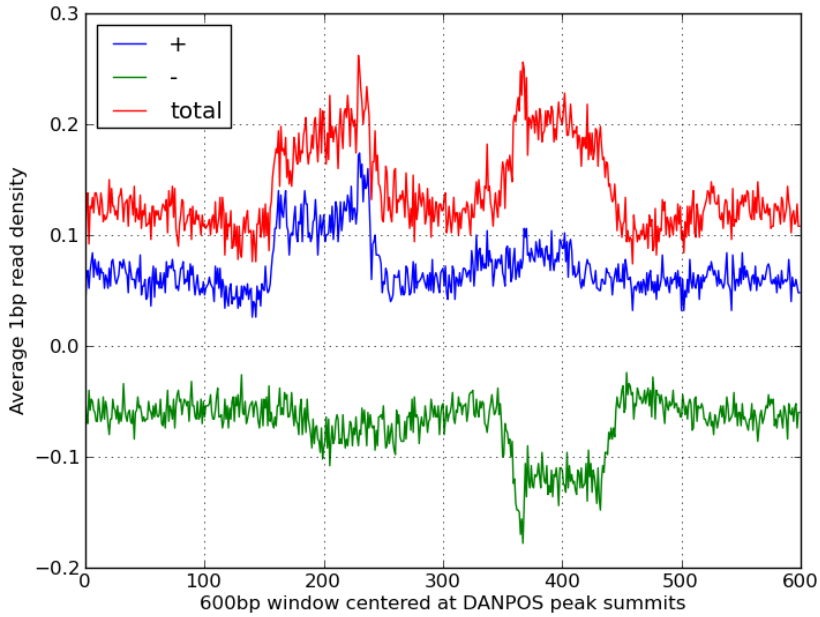


(b) Reads mapped to - strand.

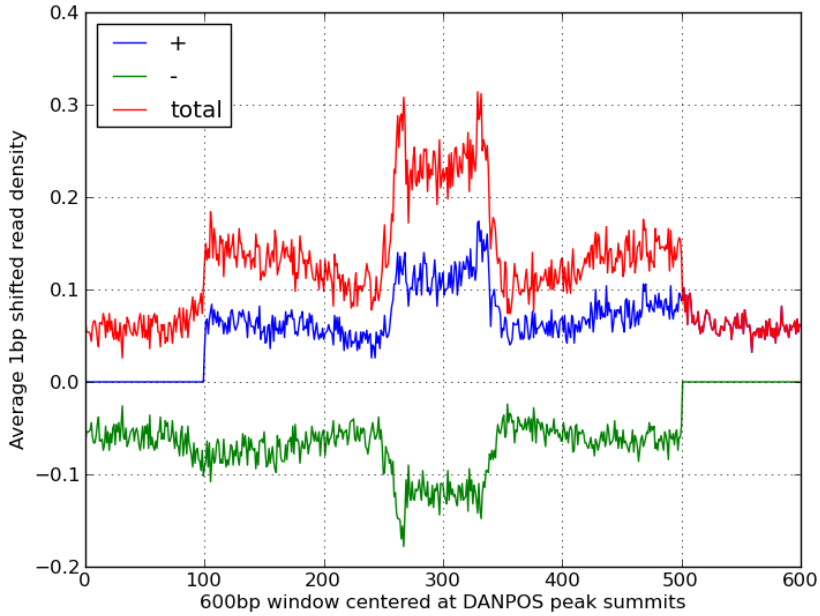


(c) Reads mapped to both strands.

Figure 2.4: Heat maps displaying the 1bp resolution read distribution in 600bp regions centered at nucleosome summit positions (most significant 500) detected by DANPOS.



(a) Fragments are unshifted.



(b) Fragments are shifted by half of the phase shift value between + strand signal and - strand signal.

Figure 2.5: Average 1bp read density profiles in 600bp regions centered at DANPOS reported nucleosome summits.

Chapter 3

Model Description

3.1 Model Description

Let $\Theta = \{\theta^1, \theta^2, ..., \theta^M\}$ be the set of M DBFs consisting of nucleosomes and $M-1$ different TFs where θ^i denotes the DBF i. $S = (s_1, ..., s_L)$ is the genome sequence of the nucleotides in the inspected region which is L bp long. For notational simplicity, L is expected to be a multiple of the *bin size* n which is 5bp by default.

Q, which stands for the number of binding sites in S, is unknown. All binding sites of any DBF θ^i are assumed to be $\lceil \frac{\delta^i}{n} \rceil \times n$ bp long where δ^i is the average motif length of DBF θ^i. $Q = c$ denotes that there are c non-overlapping binding sites which can belong to any of the M DBFs. $K = \{k_1, ..., k_c\} \subset S$ is the set of start positions of these c non-overlapping binding sites. Finally, $\pi \subset \{\theta^1, ..., \theta^M\}^c$ stands for a configuration indicating which DBFs hold which binding sites. A tuple consisting of configuration π with the corresponding set K, can be referred to as a *state*. For instance, when $c = 3$, $M = 7$ and $S = (4567, ..., 4700)$ on chromosome one, one such configuration might be $\pi = (\theta^3, \theta^7, \theta^3)$ with the corresponding start positions $K = \{k_1 = 4569, k_2 = 4615, k_3 = 4690\}$ assuming $\lceil \frac{\delta^3}{n} \rceil \times n \leq 15$ and $\lceil \frac{\delta^7}{n} \rceil \times n \leq 75$.

Read density profiles of Chip-seq and MNase-seq data, for which the analysis results were presented in Chapter 2, can be used for assessing the probability of a configuration π with a specified set K. We can denote the set of binned read density profiles of DBFs' high throughput data mapped to S as $D = \{d_1, ..., d_M\}$ where d_i stands for the binned read distribution profile of high throughput data for θ^i along the inspected region S. That is, when $x_{i,j}$ represents the number of 1bp resolution reads mapped to bin j for θ^i:

$$d_i = (x_{i,1}, x_{i,2}, ..., x_{i,\frac{L}{n}}). \tag{3.1}$$

Using Bayes' rule, the probability of the binding site positions K and configuration π, given D is:

$$P(K, \pi|D) = \frac{P(D|K, \pi)P(K, \pi)}{P(D)} \tag{3.2}$$

Since $D = \{d_1, ..., d_M\}$, when we assume conditional independence between read distribution profiles given a particular configuration and start positions Equation 3.2 can be rewritten as:

$$\begin{aligned} P(K, \pi|D) &= P(K, \pi|d_1, ..., d) \\ &= \frac{P(d_1, ..., d_M|K, \pi)P(K, \pi)}{P(d_1, ..., d_M)} \\ &= \frac{P(d_1|K, \pi)...P(d_M|K, \pi)P(K, \pi)}{P(d_1, ..., d_M)} \end{aligned} \tag{3.3}$$

The peak summits reported by MACS provide us empirical prior information about the number of possible binding sites of each TF. The prior probability of TF θ^i having b_i number of binding sites in state (K, π) is modeled by a Poisson distribution with λ_i being equal to the number of peak summits reported by MACS for the inspected region. For the nucleosomes, we lack such information and we model the prior probability distribution of the number of nucleosomes in the Lbp long region as Unif$[0, \frac{L}{150}]$. Therefore, the prior $P(K, \pi)$ is:

$$P(K, \pi) = \mathit{Poi}(b_1|\lambda_1)...\mathit{Poi}(b_{m-1}|\lambda_{M-1})\mathit{Unif}[b_m|0, \frac{L}{150}] \tag{3.4}$$

$P(d_i|K, \pi)$ denotes the marginal likelihood of the binned read density profile of DBF θ^i along S, given configuration π and the set of binding site start positions K.

Based on our observations explained in Chapter 2, we can conclude that d_i is composed of bins belonging to enriched and background regions. For TFs, the enriched regions are assumed to be fragment length long windows centered (referred to as *enriched windows* hereinafter) at the binding sites and are of equal length for all of the binding sites of a specific TF. Whereas, for nucleosomes, the middle 75bp window is considered to be the enriched region of a nucleosome site of 150bp long. When two or more binding sites of a TF are close to each other, the enriched regions around these binding sites

overlap forming *combined enriched regions* (defined later). The likelihood $P(d_i|K,\pi)$ is the product of the likelihood of the enriched bins $P(d_{i-en}|K,\pi)$ and the likelihood of the background bins $P(d_{i-bg}|K,\pi)$

$$P(d_i|K,\pi) = P(d_{i-en}|K,\pi)P(d_{i-bg}|K,\pi). \tag{3.5}$$

Note that, for each d_i, (K,π) defines the enriched and background bins. Instead of $P(d_{i-en(K,\pi)}|K,\pi)$ and $P(d_{i-bg(K,\pi)}|K,\pi)$ we write $P(d_{i-en}|K,\pi)$ and $P(d_{i-bg}|K,\pi)$ for notational brevity.

Negative binomial distribution is a popular choice for modelling background read counts [67]. We model the distribution of the background bin read counts with poisson distribution with gamma prior. The poisson and gamma distributions being conjugate, the marginal likelihood can be calculated in closed form using the negative-binomial distribution [68]. For θ^i, if we denote the set of indices corresponding to background regions at state (K,π) as $I_{i-bg}(K,\pi)$ and assume an independent gamma prior for each bin, then the the background likelihood $P(d_{i-bg}|K,\pi)$ can be calculated as the product of the likelihood of the read counts in each background bin:

$$\begin{aligned} P(d_{i-bg}|K,\pi,\lambda_{i-bg}) &= \prod_{j\epsilon I_{i-bg}(K,\pi)} P(x_{i,j}|\lambda_{i-bg}) \\ P(\lambda_{i-bg}|\alpha,\beta) &= Ga(\alpha,\beta) \end{aligned} \tag{3.6}$$

$$\begin{aligned} P(d_{i-bg}|K,\pi) &= \prod_{j\epsilon I_{i-bg}(K,\pi)} \int P(x_{i,j}|\lambda_{i-bg})P(\lambda_{i-bg}|\alpha,\beta)d\lambda_{i-bg} \\ &= \prod_{j\epsilon I_{i-bg}(K,\pi)} \int \mathrm{Poi}(x_{i,j}|\lambda_{i-bg})\mathrm{Ga}(\lambda_{i-bg}|\alpha,\beta)d\lambda_{i-bg} \\ &= \prod_{j\epsilon I_{i-bg}(K,\pi)} \mathrm{NB}(x_{i,j}|\alpha,\beta) \end{aligned} \tag{3.7}$$

We have used the empirical Bayes approach and estimated the background NB parameters by using the bin read counts in a [-2500, + 2500] interval around the inspected region. For TFs, the bin count values of all bins in this interval are used for background parameter estimation. However, considering that 75 to 90 percent of eukaryotic genomic DNA is packaged into nucleosomes and each $\sim$ 210bp (147 for the particle plus $\sim$ 60bp for the one linker region) nucleosomal region has a corresponding 75bp enriched window in our model, we have used the 85% lowest bin read counts to infer the parameters of the nucleosome background signal.

Let $T_{i,j}$ denote the total read count in enriched window j centered at a binding site of DBF θ^i . Based on our observations for distribution of the total number of reads in enriched windows centered at peak summits reported by MACS, we model the distribution of $T_{i,j}$ with poisson distribution with gamma prior. Given $T_{i,j}$, the read counts of the individual bins in the enriched window, $Y_{i,j}$, are assumed to follow the multinomial distribution with bin level probabilities Φ which describes the peak shape around a single binding site[1]. Peaks in real data vary in widths, heights, and shapes, possibly due to various biological and technical factors. Therefore, we use a dirichlet prior for Φ which provides a flexible model to allow the heterogeneity and variation in the peak shapes.

When there are b_i non-overlapping enriched windows of θ^i in state (K, π), the likelihood of the combination of the enriched regions $P(d_{i-en}|K, \pi)$ is:

$$\begin{aligned} P(d_{i-en}|K,\pi) &= \prod \left(\int P(T_{i,j}|\lambda_{en}) P(\lambda_{en}|\alpha', \beta') d\lambda_{en} \right) \cdot \\ &\quad \left(\int P(Y_{i,j}|\Phi, T_{i,j}) P(\Phi|\alpha^*) d\Phi \right) \qquad (3.8) \\ &= \prod \left(\int \mathrm{Poi}(T_{i,j}|\lambda_{en}) \mathrm{Ga}(\lambda_{en}|\alpha', \beta') d\lambda_{en} \right) \cdot \\ &\quad \left(\int \mathrm{MN}(Y_{i,j}|\Phi, T_{i,j}) \mathrm{Dir}(\Phi|\alpha^*) d\Phi \right) \qquad (3.9) \\ &= \prod_{j=1}^{b_i} \left(\mathrm{NB}(T_{i,j}|\alpha', \beta') \cdot \mathrm{Polya}(Y_{i,j}|T_{i,j}, \alpha^* \right) \qquad (3.10) \end{aligned}$$

In equation 3.10, (α', β') are parameters of the gamma distribution which is the prior for λ_{en}. α^* stands for the prior dirichlet parameters which specifies the template unimodal peak profile. Since each bin has its own specified multinomial probability value, number of the parameters in α^* is equal to the number of bins in an enriched window. (This bin count is equal to $\lceil \frac{\text{fragment length}}{\text{bin size}} \rceil$ for TFs and $\lceil \frac{75}{\text{bin size}} \rceil$ for nucleosomes.) For enriched windows containing only one binding site, to set α^*, we calculate the binned read distribution signals in windows centered at the 4000 most significant peak summits detected by MACS (for TFs) or 8000 DANPOS nucleosome peak summits. Then, we estimate α^* by finding the values which maximizes the likelihood of this array of signals. Since there is no closed-form solution for

[1]Note that $Y_{i,j}$ is a set of consecutive read counts from d_i,i.e., $Y_{i,j} = (d_{i,r}, d_{i,r+1}, ..., d_{i,r+l}) = (x_{i,r}, x_{i,r+1}, ..., x_{i,r+l})$

the maximum-likelihood estimate of Dirichlet-multinomial compound distribution (also known as multivatiate Polya distribution) parameters, we find the maximum-likelihood estimate of $\alpha^*$by using the L-BFGS-B optimization algorithm [74] implemented in Python scipy.optimize module. However, for some DBFs we have observed that the average of this signal array does not project the read distribution around individual peak summits, which is in general a unimodal peak centered at the peak summit. Therefore, we decided to use a Gaussian shaped function $\exp(-\frac{(x-z)^2}{2(2z)^2})$ for regularizing α^* while making sure that sum of the parameters in resulting α^* is equal to the sum of the parameters in maximum likelihood estimate, i.e., if the number of bins in the enriched window centered at a single binding site is $2z$ then:

$$\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_{2z}^*) \tag{3.11}$$

$$\alpha_j^* = \exp(-\frac{(j-z)^2}{8z^2}) \tag{3.12}$$

In some states where two or more binding sites of the same TF are closely located, the enriched windows, each of which is centering at a different binding site, may overlap. In these cases, the prior Dirichlet parameters for the "combined" enriched window are calculated as a combination of the dirichlet parameters of the individual enriched windows such that the corresponding parameters of the overlapped bins are summed up.

Integrating out the enriched window specific multinomial probabilities Φ, we can obtain the marginal bin counts distribution conditional on total count of the enriched window $T_{i,j}$. This marginalization is computed in closed form by Polya distribution.

3.2 Monte Carlo Markov Chain Estimation

In addition to the fact that calculating the necessary normalization factor $P(D)$ is extremely difficult, computing the posterior $P(K, \pi|D)$ for all K and π is impossible. However, we can draw samples from the target distribution by using Markov Chain Monte Carlo (MCMC) [69] framework which is scaled well with the dimensionality of the sample space. For this purpose we use a Metropolis-Hasting algorithm by which we construct a Markov chain such that its unique stationary probability distribution $P((K, \pi))$ converges to our target distribution $P(K, \pi|D)$ irrespective of the choice of the initial distribution.

A distribution is said to be stationary, with respect to Markov chain if each step in the chain leaves the distribution stationary [69]. That is,

with transition probabilities $T((K', \pi'), (K, \pi))$, the distribution $P((K, \pi))$ is stationary if :

$$P((K', \pi')) = \sum_{((K,\pi))} T((K', \pi'), (K, \pi))P((K, \pi))$$

In order to guarantee an acceptable approximation of the simulated model, the Markov chain should have a unique stationary distribution. It is guaranteed that $P((K, \pi))$ is a unique stationary distribution when the following two conditions are met:

- **existence of stationary distribution:** Detailed balance condition is a sufficient but not necessary way for ensuring that there exists a stationary distribution $P((K, \pi))$. A stationary distribution requires that each transition $T((K', \pi'), (K, \pi))$ is reversible. In other words, for every pair of states (K, π), (K', π'), the probability of being in state (K, π) and transit to the state (K', π') must be equal to the probability of being in state (K', π') and transit to the state (K, π):

$$P((K, \pi))T((K', \pi'), (K, \pi)) = P((K', \pi'))T((K, \pi), (K', \pi'))$$

This preliminary constraint on $T((K', \pi'), (K, \pi))$ is called irreducibility. Irreducibility property ensures that for any state of the Markov chain, there is a positive probability of visiting all other states. That brings transition kernel T in allowing for moves all over the state-space. In the discrete case that means no matter what the starting state is, the Markov chain has a positive probability of eventually reaching any region of the state space.

- **uniqueness of stationary distribution:** This is guaranteed by ergodicity property, which requires that every state must be:

  - aperiodic: This property ensures that the system does not return to the same state at fixed intervals.
  - positive recurrent: This property ensures that the system will return to the same state with nonzero probability and expected recurrence time is finite.

  The notion of positive recurrency is necessarily satisfied for irreducible chains on a finite space [69]. Therefore, it can be shown that a finite state irreducible Markov chain is ergodic if it has an aperiodic state. In other words, a model is said to hold the ergodic property if there is a finite number N such that any state can be reached from any other state in exactly N steps. When ergodicity property holds, the chain

will converge to the same stationary distribution $P((K, \pi))$ no matter what the initial state is.

Based on these requirements, we define the proposal distribution $G((K', \pi')|(K, \pi))$, which proposes new state (K', π') based on the current state (K, π), as follows:

- **Move 1:** with probability p_1, propose a new non-occupied and non-overlapping binding site to a DBF which is randomly chosen among DBFs that are eligible for having an additional non-overlapping binding site in the current state. To increase acceptance rate, proposed binding site is selected by the use of a multinomial probability distribution which always has non-zero set of parameters that are directly proportional to the chosen DBF's binned read density profile at the available non-occupied, non-overlapping binding sites.

- **Move 2:** with probability p_2, delete a randomly chosen binding site of a randomly chosen DBF which has a binding site in the current configuration. Chosen binding site is selected by the use of a multinomial probability distribution which always has non-zero set of parameters that are directly proportional to the binned read density profile of the DBF at its current binding sites.

- **Move 3:** with probability p_3, shift a randomly chosen binding site of a DBF by 1 bin to the left provided that a binding site of this DBF exists and the shifted position does not overlap with any other binding sites

- **Move 4:** with probability p_4, shift a randomly chosen binding site of a DBF by 1 bin to the right provided that a binding site of this DBF exists and the shifted position does not overlap with any other binding sites

- **Move 5:** with probability p_5, swap randomly chosen binding sites of two TFs with equal motif lengths which have at least one binding site in the current configuration

where

$$p_1 > 0,\ p_2 > 0,\ p_3 > 0,\ p_4 > 0, p_5 > 0$$

$$p_1 + p_2 + p_3 + p_4 + p_5 = 1$$

The candidate sample is accepted with probability:

$$
\begin{aligned}
A((K',\pi'),(K,\pi)) &= \min\left\{1, \frac{P(K',\pi'|D)}{P(K,\pi|D)} \times \frac{G((K,\pi)|(K',\pi'))}{G((K',\pi')|(K,\pi))}\right\} \\
&= \min\left\{1, \frac{P(d_1|K',\pi')...P(d_M|K',\pi')P(K',\pi')}{P(d_1|K,\pi)...P(d_M|K,\pi)P(K,\pi)}\right\} \quad (3.13) \\
&\quad \left\{\frac{G((K,\pi)|(K',\pi'))}{G((K',\pi')|(K,\pi))}\right\} \quad (3.14)
\end{aligned}
$$

If the candidate sample is accepted at time t, then the next state is (K',π') at time $t+1$, otherwise the candidate sample (K',π') is discarded and the state at $t+1$ is set to (K,π) and another sample is drawn from the distribution $G((K',\pi')|(K,\pi))$. That is, when a candidate sample is rejected, the previous sample is included instead in the final list of samples, leading to multiple copies of the samples.

The resulting chain satisfies the irreducibility and aperiodicity constraints which are required for a unique stationary distribution. Since the inspected region is finite, the number of all possible DBF binding configurations is finite. We assume that each one step move has a non-zero probability which makes any state (K,π) be reached from any other state (K',π') by following a finite number of moves proposed by G. Therefore, irreducibility of the chain holds. Aperiodicity of our chain is also assured because in each step, there is a positive probability of choosing the reverse moves of the moves that have been made so far which makes the probability of moving from (K,π) back to (K,π) in two or more number of steps non-zero. For instance at state (K,π) the chain has a non-zero probability of moving to state (K',π') and stay at this state for one or more steps and return back to state (K,π).

3.3 Probabilities of Forward and Reverse moves

In a region consisting of $r = \frac{L}{n}$ number of bins, $B = (b_1,...,b_M)$ vector contains the number of non-overlapping binding sites for each of the M DBFs in this region in the current state. For simplicity, lets suppose each DBF has a motif length equal to the bin size, and thus b_i is equal to the number of bins occupied by the binding sites of DBF θ^i. When computing the probability in Equation 3.13, $G((K',\pi')|(K,\pi))$ and $G((K,\pi)|(K',\pi'))$ can be calculated for 5 possible moves of the proposal distribution as follows:

Forward and Reverse Move Probabilities of Move 1:

In this simplified case where each DBF has a motif length equal to the bin size, the expression $r - \sum_{j=1}^{M} b_j$ is equal to the number of candidate

binding sites in the current state. The probability of a candidate binding site to be selected is directly proportional to the binned read count of the selected DBF at this location. By the use of the multinomial distribution, candidate non-occupied, non-overlapping binding site at k^{th} bin is selected with non-zero probability $\frac{x_{ik}+2}{\sum_s(x_{is}+2)}$ where $x_{i,s}$ represents the number of 1bp resolution reads mapped to bin s for θ^i. Therefore, the forward and reverse move probabilities of addition move are:

$$G((K',\pi')|(K,\pi)) = \frac{p_1}{m_1}\left(\frac{x_{ik}+2}{\sum_s(x_{is}+2)}\right) \tag{3.15}$$

$$G((K,\pi)|(K',\pi')) = \frac{p_2}{m_2^{'}}\left(\frac{x_{ik}+2}{\sum_s^{b_z'}(x_{is}+2)}\right) \tag{3.16}$$

Here m_1 represents the number of DBFs which are eligible for having an additional binding site in the current state. Therefore, it is equal to the number of DBFs whose motif length is smaller or equal to any of the available free spaces in the current binding configuration. When motif lengths of DBFs are some multiples of the bin size, the number of candidate sites should be calculated accordingly. $m_2^{'}$ is the number of DBFs which has at least one binding site in the proposed state and $b_z^{'}$ is the number of binding sites of the selected DBF in the proposed state.

Forward and Reverse Move Probabilities for Move 2:

$$G((K',\pi')|(K,\pi)) = \frac{p_2}{m_2}\left(\frac{x_{ik}+2}{\sum_s^{b_z}(x_{is}+2)}\right) \tag{3.17}$$

$$G((K,\pi)|(K',\pi')) = \frac{p_1}{m_1^{'}}\left(\frac{x_{ik}+2}{\sum_s^{r-\sum_{j=1}^{m} b_j'}(x_{is}+2)}\right) \tag{3.18}$$

Forward and Reverse Move Probabilities for Move 3:

$$G((K',\pi')|(K,\pi)) = \frac{p_3}{m_3}\frac{1}{b_z^*} \tag{3.19}$$

$$G((K,\pi)|(K',\pi')) = \frac{p_4}{m_4^{'}}\frac{1}{b_z^{\sim'}} \tag{3.20}$$

Forward and Reverse Move Probabilities for Move 4:

$$G((K',\pi')|(K,\pi)) = \frac{p_4}{m_4}\frac{1}{b_z^{\sim}} \tag{3.21}$$

$$G((K,\pi)|(K',\pi')) = \frac{p_3}{m_3^{'}}\frac{1}{b_z^{*'}} \tag{3.22}$$

In equations 3.19-3.22, m_3 and $m_3^{'}$ are the number of DBFs which have at least one binding site eligible for a left shift in current and next state respectively. Similarly, m_4 and $m_4^{'}$ are the number of DBFs which have at least one binding site eligible for a right shift in current and next state. In these equations, b_z^* and $b_z^{*'}$ denote the number of binding sites of the selected DBF which can be shifted to left in current and proposed states respectively. Likewise, $b_z^{\sim}$ and $b_z^{\sim'}$ denote the number of binding sites of the selected DBF which can be shifted to right in current and proposed states.

Forward and Reverse Move Probabilities for Move 5:

Forward and reverse move probabilities of move 5 are always equal to each other, which can be computed by the following function:

$$G((K',\pi')|(K,\pi)) = G((K,\pi)|(K',\pi')) = \frac{p_5}{m_2(m_2-1)}\frac{1}{b_x}\frac{1}{b_y} = \frac{p_5}{m_2^{'}(m_2^{'}-1)}\frac{1}{b_x^{'}}\frac{1}{b_y^{'}} \tag{3.23}$$

Again m_2 and $m_2^{'}$ are the number of DBFs which has at least one binding site in the current and proposed state respectively. Since the total number of occupied binding sites do not change with this move, m_2 is always equal to $m_2^{'}$. Likewise, $b_x$and b_y are the number of non-overlapping binding sites of θ^x and θ^y which do not change in the proposed state.

The sequence of samples obtained by the employment of our proposal distribution is used to approximate the posterior distribution $P(K,\pi|D)$. Each sample defines a configuration π with the corresponding start positions K which together specify the binding sites of the M different DBFs and the sites which are not bound by any of the DBFs. When all of the collected samples are taken into account, an $M \times \frac{L}{n}$ probability matrix A is calculated where each element a_{ij} indicates the binding probability of θ^i to the sites in the j^{th} bin of the considered region. Since these probability values are computed by the use if the MNase-seq and ChIP-seq data, they automatically reflect the sequence preferences and cell concentrations of the DBFs as well as the effects of the competitive binding and other factors affecting DBFs' binding locations and binding frequencies.

3.4 Convergence Monitoring

Even though the chain is ergodic as shown above, it is important to monitor the convergence of the algorithm and make sure that the distribution of the

chain does not change over time. For this purpose, we periodically find the maximum of the absolute difference between the posterior distributions of the samples obtained by two separate chains which are run simultaneously and stop the chains when this distance is below a user defined maximum threshold.

Chapter 4

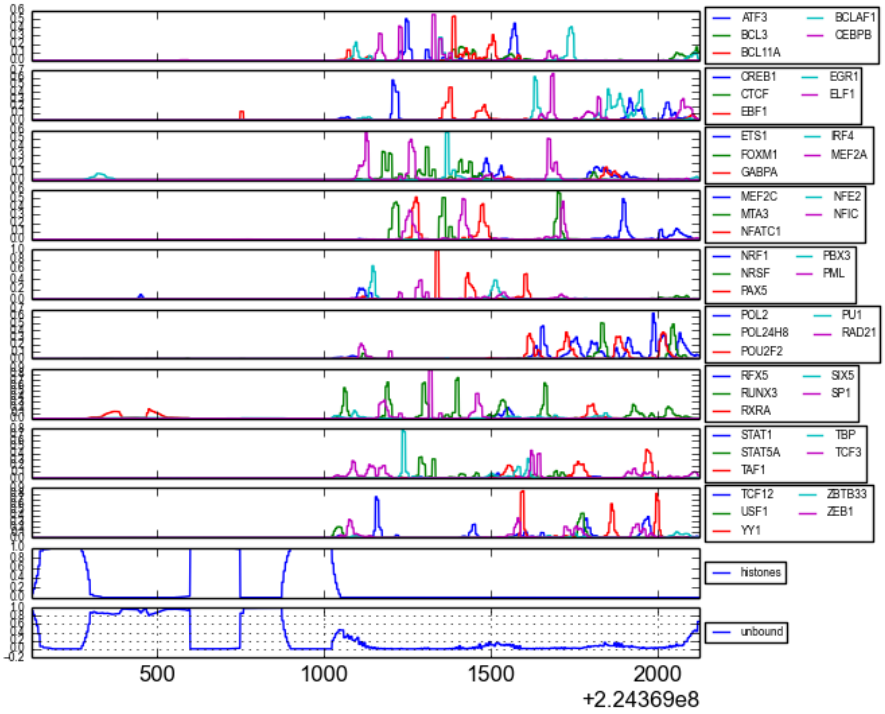
Experiment Results

We applied our model to 46 ChIP-seq data and MNase-seq data at 300 [-1000, +1000] regions around transcription start sites (TSSs) on chromosome one. Names of the TFs are listed in the Materials section. These regions are selected according to their MACS reported peaks abundance.

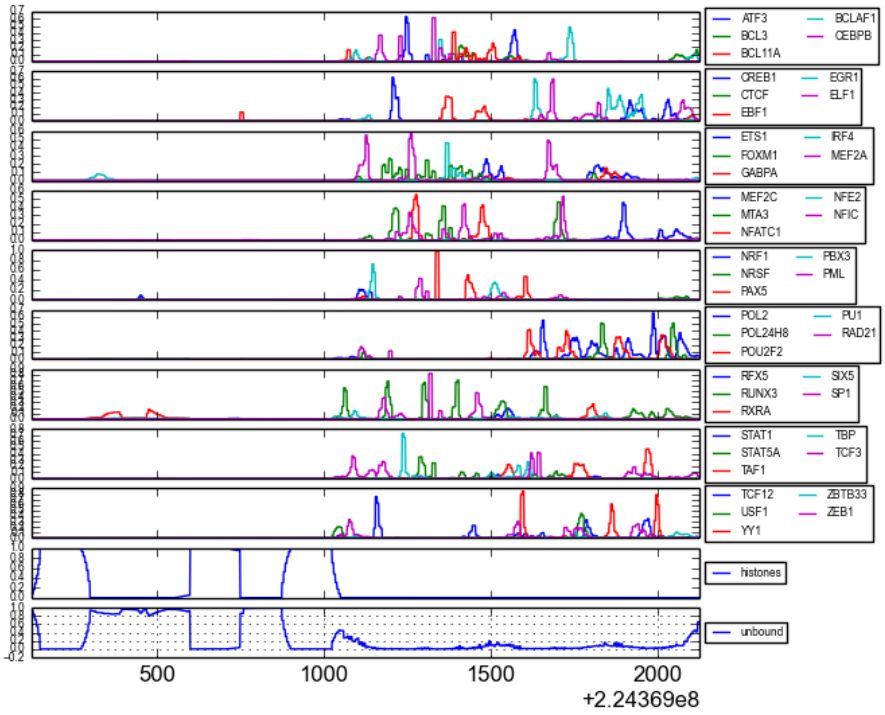
During our simulations, the convergence threshold for the maximum of the absolute difference between the posterior distributions of the two separate chains is set to be between $0.08 - 0.1$. In the burn-in period, the first $200,000$ samples are thrown away. The maximum number of acquired samples is set to be $26,000,000$, therefore if the threshold is not reached in the specified running time, the posteriors of the chains are calculated based on this number of accepted samples.

To illustrate a possible outcome of our proposed method, Figure 4.1 displays chain one and chain two posterior binding probabilities of 45 TFs and nucleosomes in a [-1000, +1000] interval around a TSS. In each of the subplots, the first nine tracks show the probability values for the TFs, track ten is for the nucleosomes and track eleven stands for the probability values of each 5bp bin being not occupied by any of the considered DBFs.

The two chains are run as independent processes and the maximum of the absolute difference between the chains is checked every 500 seconds by the main process. At each of these controls, if the threshold value is reached, the chains are stopped and the output posterior probability distribution is calculated by taking the mean of the two chains' posteriors. Otherwise, the algorithm runs until both of the chains accumulate at least $26,000,000$ accepted samples.



(a) Posterior probabilities of chain one.



(b) Posterior probabilities of chain two.

Figure 4.1: Posterior probabilities for a [-1000, +1000] interval around a TSS.

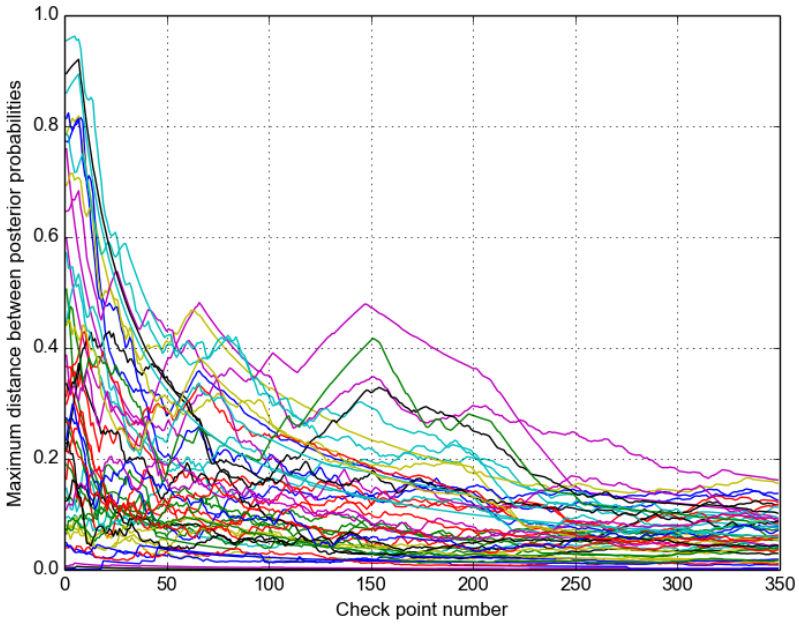
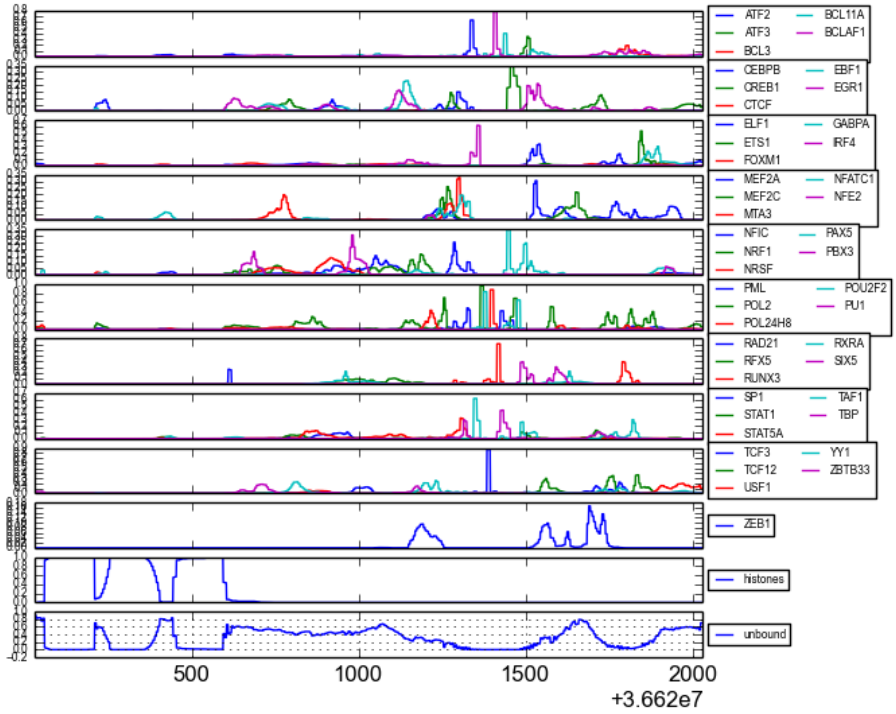


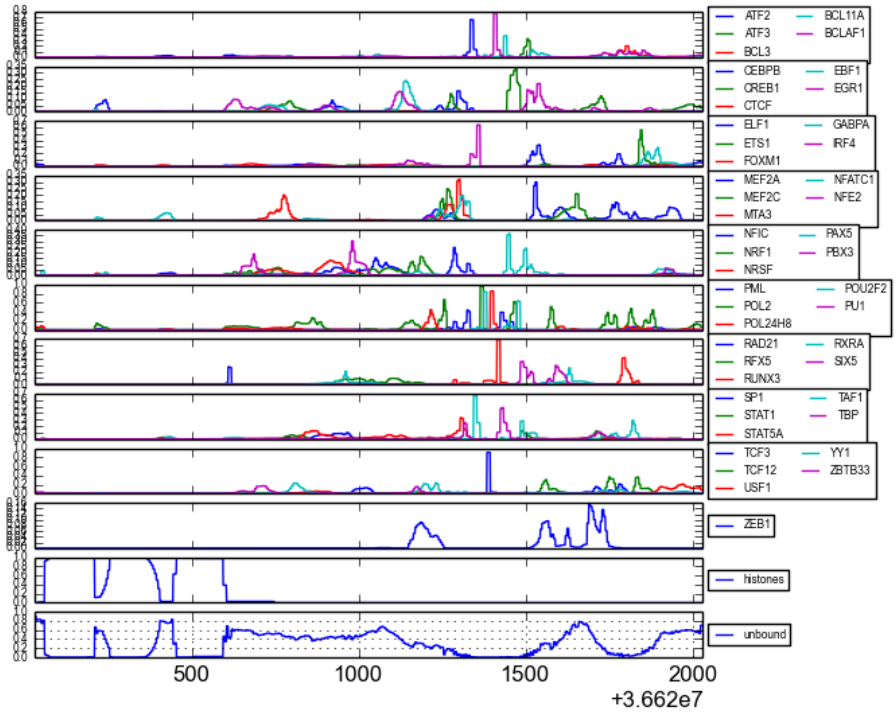
Figure 4.2: Time course maximum chain posterior difference values of 46 DBFs.

The probability values shown in Figure 4.1 are obtained in a simulation where the convergence threshold was not reached at the end of 349 checks and the acquired $26,000,000$ samples were used to estimate the posteriors. Figure 4.2 displays the evolution of the maximum posterior difference values of 46 DBFs through these 349 time course check points. During our simulations, we generally observed this kind of exponential decrease in the distance values between the chains.

As another example, Figure 4.3 displays the posterior probability values of the two chains in an experiment where the convergence was reached with a threshold of 0.1. Figure 4.4 shows the convergence monitoring of this experiment in which all 47 DBFs were utilized.



(a) Posterior probabilities of chain one.



(b) Posterior probabilities of chain two.

Figure 4.3: Posterior probabilities for a [-1000, +1000] interval around a TSS.

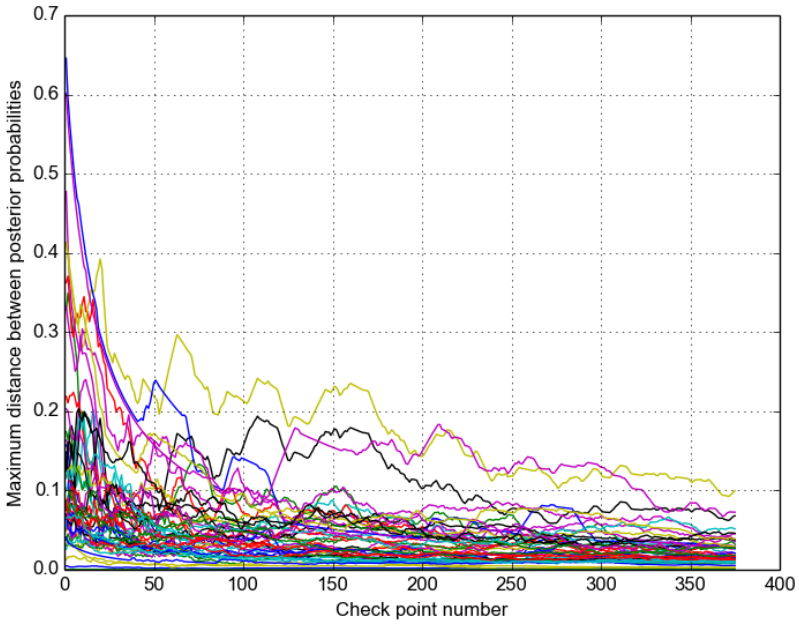


Figure 4.4: Time course maximum chain posterior difference values of 47 DBFs.

4.1 Performance Evaluation for Accurate Binding Site Detection

In order to determine the accuracy of our proposed model in detecting the binding sites of the TFs, we compared the peaks of our posterior values with MACS reported results based on the annotated TF motif sites. During these comparisons, for each TF we identified as many highest values out of the posterior probabilities as the number of peaks reported by MACS and referred the bin locations of these maximum posterior values as our predicted TF binding sites.

We used the 'factorbookMotifPos' and 'factorbookMotifCanonical' tables of the hg19 database of the UCSC bioinformatics site to obtain the positions of the annotated TF motif sites. Then, we measured the distance between the motif sites and the midpoints of our predicted binding sites, and the distance between the MACS peak summits and the motif sites. Since, some of the peaks might have been due to indirect binding to the DNA, we only considered the distance values which were below 200bp. We calculated all such distance values for 46 TFs in 300 [-1000, +1000] regions around TSSs. As

an example, Figures 4.6 and 4.7 display the histograms of the distance values between the predicted binding locations and ELF1, YY1 motif sites in the inspected 300 regions respectively. For these TFs, we see that in general our maximum posterior values are closer to the motif sites than MACS reported binding sites. Figure 4.5 displays the comparison of the means of these distance values for all of the considered TFs which has at least two distance measurements (smaller than 200bp) in total. With these comparisons we again observe that, our predictions are usually closer to the motif sites than the MACS reported peak summits.

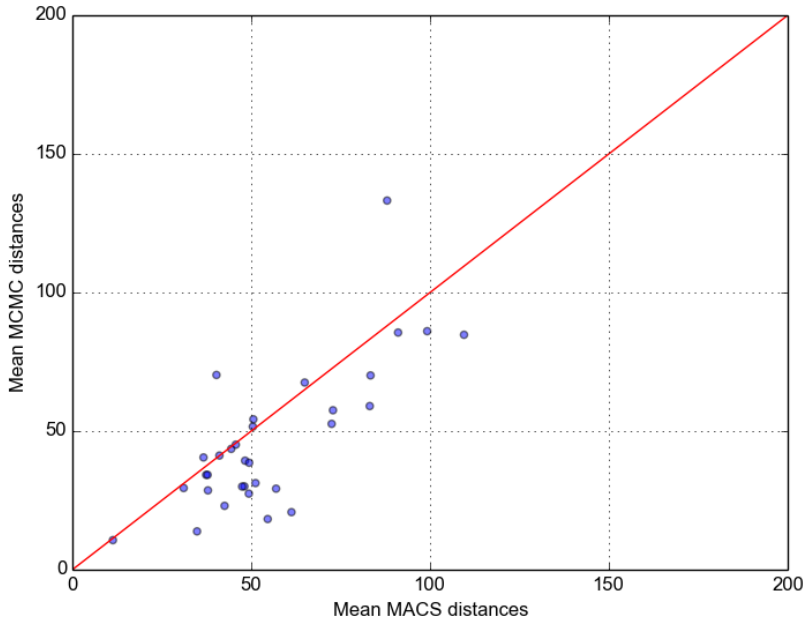
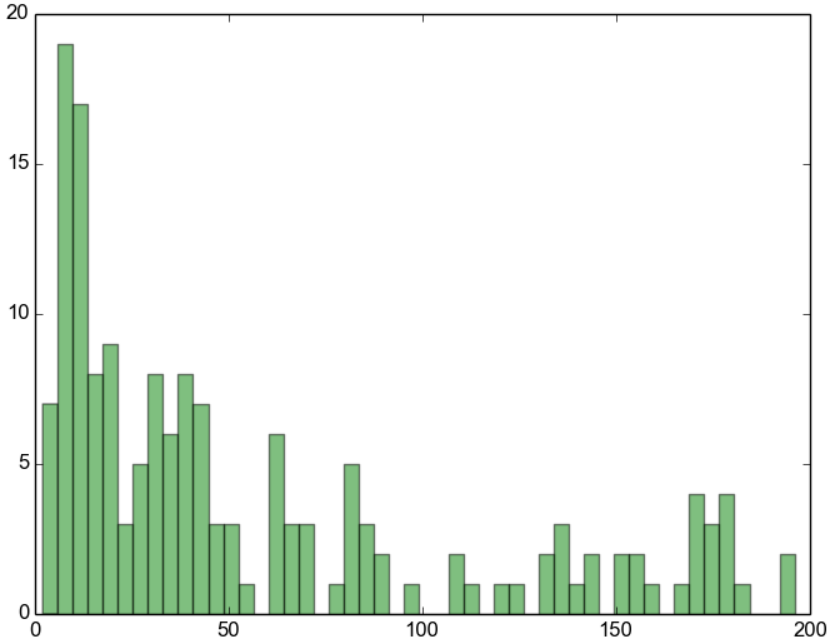
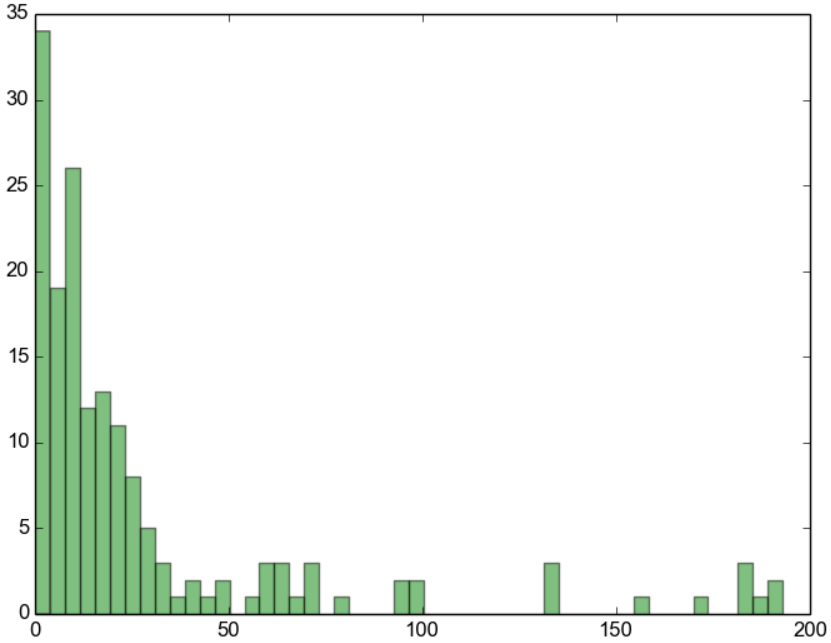


Figure 4.5: Mean distance values for each TF between MACS reported binding sites-motif sites versus our predicted binding sites-motif sites.

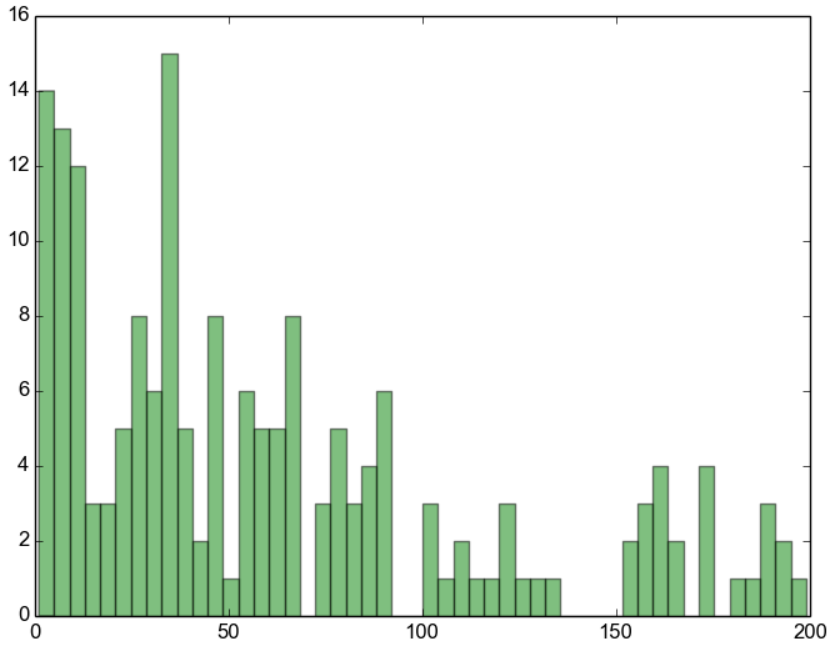


(a) Histogram for the distance values between the MACS ELF1 peak summits and annotated ELF1 motif sites.

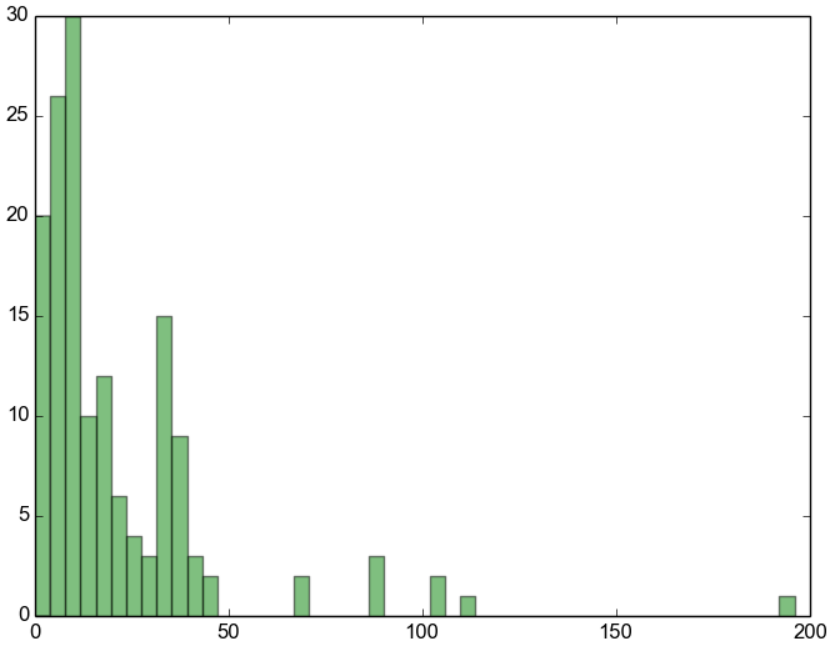


(b) Histogram for the distance values between our method's predicted binding sites and annotated ELF1 motif sites.

Figure 4.6: Histograms for the distance values, which are smaller than 200bp, between ELF1 motif sites and the binding site predictions.



(a) Histogram for the distance values between the MACS YY1 peak summits and annotated YY1 motif sites.



(b) Histogram for the distance values between our method's predicted binding sites and annotated YY1 motif sites.

Figure 4.7: Histograms for the distance values, which are smaller than 200bp, between YY1 motif sites and the binding site predictions.

Chapter 5

Discussion

In this study, we proposed a probabilistic model for the organization of DBFs on DNA using ChIP-seq and MNase-seq data. Our model takes into account the combinatorial nature of binding probabilities of DBFs to DNA sites, and annotates the binding sites using a probability interval. Moreover, it includes the competition between different factors, which, according to our results, increase the precision of binding site estimates. Unlike thermodynamic models, which also aim to model the combinatorial binding pattern of the DBFs, our model does not assume thermodynamic equilibrium conditions and intrinsically considers in vivo conditions by the use of the high-throughput ChIP-seq and MNase-seq data. Finally, our proposed method is applicable on high number of DBFs which is an important advantage for studies that target integrated analysis of DBFs.

In our experiments, we have used high-throughput ChIP-seq data for 46 TFs and MNase-seq data for nucleosomes and analyzed 300 $[-1000, +1000]$ regions centered at TSSs. Throughout the experiments, we have observed efficient convergence statistics of the MCMC chains. In order to evaluate our binding site predictions, we have compared the distances between the maximums of the posterior probability distributions and MACS reported peaks with respect to the annotated motif sites. We have seen that for most of the TFs, our predictions have smaller mean distance values to the motif sites.

Our method's bottleneck is the slower mixing time to reach stationary distribution, which specifically shows up during integration of high number of DBFs. Therefore, it is difficult to apply it to large genomic regions. That being said, we claim that it can be especially useful for the analysis of cis-regulatory regions like promoters and enhancers.

Bibliography

[1] V.B. Teif and K. Rippe, Calculating transcription factor binding maps for chromatin. In Proceedings of Briefings in Bioinformatics. 2012, 187-201.

[2] T. Wasson and A. Hartemink, An ensemble model of competitive multi-factor binding of the genome. Genome Res. 2009 19 (11): 2101-2112

[3] J. Gertz, ED Siggia, BA Cohen, Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature 2009;457:215-8

[4] X. He, MA. Samee, C. Blatti, S. Sinha, Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput Biol 6: e1000935. 2010, doi:10.1371/journal.pcbi.1000935.

[5] Tali Raveh-Sadka, Michal Levo, Eran Segal, Incorporating nucleosomes into thermodynamic models of transcription regulation. Genome Res. 2009 August; 19(8): 1480–1496. doi: 10.1101/gr.088260.108

[6] T. Kaplan, X.Y. Li, P.J. Sabo, S. Thomas, J.A. Stamatoyannopoulos, M.D. Biggin, M.B. Eisen, Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genet. 2011 Feb 3;7(2):e1001290. doi: 10.1371/journal.pgen.1001290.

[7] B.R. Cairns, The logic of chromatin architecture and remodelling at promoters. Nature 461, 2009, 193-198 | doi:10.1038/nature08450

[8] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, U. Gaul, Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature. 2008;451(7178):535-40. doi: 10.1038/nature06496

[9] K.E. Van Holde, Chromatin. 1989, New York,Springer.

[10] Karolin Luger, et al. Crystal structure of the nucleosome core particle at 2.8 Å resolution, Nature 389.6648, 1997, 251-260.

[11] A.H. Brivanlou, J.E. Darnell, Signal transduction and the control of gene expression. 2002 Science 295 (5556): 813–8. Bibcode:2002Sci...295..813B. doi:10.1126/science.1066355

[12] N. Kaplan, I.K. Moore, Y. Fondufe-Mittendorf, A.J. Gossett, D. Tillo, et al., The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458: 362–366, 2009

[13] A. Mathelier, W.W. Wasserman, The Next Generation of Transcription Factor Binding Site Prediction, PLoS Computational Biology;2013, Vol. 9 Issue 9, p1

[14] G.D. Stormo, Modeling the specificity of protein-dna interactions. Quantitative Biology 2013, Vol:1: 115–130. doi: 10.1007/s40484-013-0012-4

[15] A. Stark, M.F. Lin, P. Kheradpour, J.S. Pedersen , L. Parts, et al., Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450, 2007, 219–32. doi: 10.1038/nature06340

[16] V. Bernard, A. Lecharny, V. Brunaud, Improved detection of motifs with preferential location in promoters. Genome 53, 2010, 739–52. doi: 10.1139/g10-042

[17] A. Arvey, P. Agius, W.S. Noble, C. Leslie, Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome research 22, 2012, 1723–34. doi: 10.1101/gr.127712.111

[18] S.J. Ho Sui, J.R. Mortimer, D.J. Arenillas, J. Brumm, C.J. Walsh, et al., oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic acids research 33, 2005, 3154–64. doi: 10.1093/nar/gki624

[19] S.J. Ho Sui, D.L. Fulton, D.J. Arenillas, A.T. Kwon, W.W. Wasserman oPOSSUM: integrated tools for analysis of regulatory motif over-representation. Nucleic acids research 35: W245–52. doi: 10.1093/nar/gkm427

[20] Th. Lin, P. Ray, G.K. Sandve, S. Uguroglu, E.P.Xing, BayCis : A Bayesian Hierarchical HMM for Cis-Regulatory Module Decoding in Metazoan Genomes. In: Vingron M, Wong L, editors, RECOMB'08 Proceedings of the 12th annual international conference on Research in

computational molecular biology. 2008, Springer Berlin Heidelberg, pp. 66–81.

[21] L. Levkovitz, N. Yosef, M.C. Gershengorn, E. Ruppin, R. Sharan, et al., A Novel HMM-Based Method for Detecting Enriched Transcription Factor Binding Sites Reveals RUNX3 as a Potential Target in Pancreatic Cancer Biology. PloS one 5, 2010, e14423. doi: 10.1371/journal.pone.0014423

[22] R.A. Salama, D.J. Stekel, Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction. Nucleic acids research 38, 2010, e135. doi: 10.1093/nar/gkq274

[23] P. Mehta, D.J. Schwab, A.M. Sengupta, Statistical Mechanics of Transcription-Factor Binding Site Discovery Using Hidden Markov Models. Journal of statistical physics 142, 2011, 1187–1205. doi: 10.1007/s10955-010-0102-x

[24] V.D. Marinescu, I.S. Kohane, A. Riva, MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC bioinformatics 6, 2005, 79.

[25] R. Durbin, S. Edddy, A. Krogh, G. Mitchison, Biological sequence analysis Probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press, 1998

[26] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, et al., Genome-wide location and function of DNA binding proteins. Science 290, 2000, 2306–2309.

[27] K.D. MacIsaac, T. Wang, D.B. Gordon, D.K. Gifford, G.D. Stormo, E. Fraenkel., An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 7, 2006, 113. doi: 10.1186/1471-2105-7-113.

[28] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-wide mapping of in vivo protein–DNA interactions. Science 316, 2007, 1497–1502.

[29] M.L. Bulyk, X. Huang, Y. Choo, G.M. Church, Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc Natl Acad Sci 98, 2001, 7158–7163.

[30] By Jkwchui [CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0) or GFDL (http://www.gnu.org/copyleft/fdl.html)], via Wikimedia Commons

[31] O. Aparicio, J.V. Geisberg, K. Struhl, Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Current Protocols in Cell Biology (University of Southern California, Los Angeles, California, USA.: John Wiley & Sons, Inc.), 2004, Chapter 17 (2004): Unit 17.7. doi:10.1002/0471143030.cb1707s23. ISBN 0-471-14303-0. ISSN 1934-2616. PMID 18228445

[32] The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. Nature , 489:57–74, 2012

[33] Elizabeth G. Wilbanks and Marc T. Facciotti. Evaluation of algorithm performance in ChIP-Seq peak detection., PloS one 5.7, 2010,: e11471.

[34] Anton Polishko, et al.,NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model., Bioinformatics 28.12 ,2012, i242-i249.

[35] http://beyondthedish.wordpress.com/tag/chromatin-fiber/

[36] Noam Kaplan, et al., Contribution of histone sequence preferences to nucleosome organization: proposed definitions and methodology., Genome Biol 11.11, 2010, 140.

[37] K. Chen, Y. Xi, X. Pan, Z. Li, K. Kaestner, J. Tyler, S. Dent, X. He, and W. Li., DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing., Genome research 23, no. 2, 2013, 341-351.

[38] A. Jansen, and Kevin J. Verstrepen, Nucleosome positioning in Saccharomyces cerevisiae., Microbiology and Molecular Biology Reviews 75, 2011, no. 2: 301-320.

[39] H.H. He, C.A. Meyer, H. Shin, S.T. Bailey, G. Wei, Q. Wang, Y. Zhang, K. Xu, M. Ni, M. Lupien, et al. 2010. Nucleosome dynamics define transcriptional enhancers. Nat Genet 42, 2010, 343–347.

[40] D.E. Schones, K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. Cell 132, 2008, 887–898.

[41] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J. Z. Wang, and J. Widom., A genomic code for nucleosome positioning., Nature 442, 2006, no. 7104 : 772-778.

[42] Heather 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 17, 2007, no. 8: 1170-1177.

[43] Guo-Cheng Yuan, and Jun S. Liu., Genomic sequence is highly predictive of local nucleosome depletion., PLoS computational biology 4, 2008, no. 1: e13.

[44] William 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 39, 2007, no. 10 : 1235-1244.

[45] H. R. Drew, and A. A. Travers. DNA bending and its relation to nucleosome positioning. J. Mol. Biol. 186, 1985, 773–790.

[46] P. T. Lowary, and J. Widom. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J. Mol. Biol. 276, 1998, 19–42.

[47] http://compbio.pbworks.com/w/page/16252888/Epigenetic%20Regulation

[48] Dustin E. Schones, and Z. Keji., Genome-wide approaches to studying chromatin modifications., Nature Reviews Genetics 9, 2008, no. 3: 179-191.

[49] Yair Field, N. Kaplan, Y. Fondufe-Mittendorf, I. K. Moore, E. Sharon, Y. Lubling, J. Widom, and E. Segal., Distinct modes of regulation by chromatin encoded through nucleosome positioning signals., PLoS computational biology 4, 2008, no. 11: e1000216.

[50] W. Hörz, and W. Altenburger., Sequence specific cleavage of DNA by micrococcal nuclease. Nucleic acids research 9, 1981, no. 12: 2643-2658.

[51] James Allan, R. M. Fraser, T. Owen-Hughes, and D. Keszenman-Pereyra., Micrococcal nuclease does not substantially bias nucleosome mapping., Journal of molecular biology 417, 2012, no. 3: 152-164.

[52] Ho-Ryun 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 5, 2010, no. 12 : e15754.

[53] Xuekui Zhang, G. Robertson, S. Woo, B. G. Hoffman, and R. Gottardo, Probabilistic inference for nucleosome positioning with MNase-based or sonicated short-read data, PloS one 7, 2012, no. 2 : e32095.

[54] Robert Schöpflin, V. B. Teif, O. Müller, C. Weinberg, K. Rippe, and G. Wedemann, Modeling nucleosome position distributions from experimental nucleosome positioning maps, Bioinformatics 29, 2013, no. 19: 2380-2386.

[55] Yong Zhang, H. Shin, J. S. Song, Y. Lei, and X. Shirley Liu, Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq, BMC genomics 9, 2008, no. 1 : 537

[56] Oscar Flores, and Modesto Orozco, nucleR: a package for non-parametric nucleosome positioning, Bioinformatics 27, 2011, no. 15: 2149-2150.

[57] Vladimir B Teif, and K. Bohinc, Condensed DNA: condensing the concepts, Progress in biophysics and molecular biology 105, 2011, no. 3 : 208-222.

[58] http://www.genome.gov/encode/

[59] http://www.genome.gov/26524238

[60] Hui Liu, R. Zhang, W. Xiong, J. Guan, Z. Zhuang, and S. Zhou, A comparative evaluation on prediction methods of nucleosome positioning., Briefings in bioinformatics, 2013, bbt062.

[61] http://genome.ucsc.edu/cgi-bin/hgFileUi?g=wgEncodeHaibTfbs

[62] S. Andrews, FastQC: A quality control tool for high throughput sequence data., 2010, [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/].

[63] Marcel Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet. journal 17, 2011, no. 1: pp-10.

[64] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol 10, 2009, no. 3 : R25.

[65] Heng Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25, 2009, no. 16 : 2078-2079.

[66] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum et al, Model-based analysis of ChIP-Seq (MACS)., 2008, Genome Biol 9, no. 9: R137.

[67] Hao Wu, and Hongkai Ji, PolyaPeak: Detecting Transcription Factor Binding Sites from ChIP-Seq Using Peak Shape Information, PloS one 9, 2014, no. 3: e89694

[68] Andrew Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. Bayesian data analysis, 2013, CRC press.

[69] Christian P. Robert, and George Casella, Monte Carlo statistical methods, 1999, Springer.

[70] Michiel Hazewinkel, Kolmogorov-Smirnov test, 2001, Encyclopedia of Mathematics, Springer, ISBN: 978-1.

[71] H. Lähdesmaki et. al., Probabilistic Inference of Transcription Factor Binding from Multiple Data Sources, PlosOne, 2008.

[72] Ranjith Padinhateeri, and John F. Marko. Nucleosome positioning in a model of active chromatin remodeling enzymes, Proceedings of the National Academy of Sciences 108, 2011, no. 19 : 7799-7803.

[73] Terrence S Furey, ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions, 2012, Nature Reviews Genetics 13, no. 12 : 840-852.

[74] Richard H. Byrd, P. Lu, J. Nocedal, and C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM Journal on Scientific Computing 16, 1995, no. 5: 1190-1208.