Modeling Bias in DNase-seq Data for Improved Chromatin Occupancy Prediction

Brian Bullins
Department of Computer Science, Duke University

17 April 2014

Contents

Abstract

Whether or not a single gene is transcribed relies on a myriad of stochastic factors which may not be adequately described by the cell's genome alone. Understanding the connection between the occupancy of a cell's chromatin and the transcription of its genes would provide insight into the dynamic regulatory dependencies that control its internal transcription state, and so enhanced techniques for modeling chromatin state would be advantageous. In this thesis we consider improved methods of integrating DNase-seq data as input for a statistical inference model which outputs a nucleotide-resolution probability distribution of the genome's chromatin occupancy profile. In particular, we focus on some initial observations made concerning the probabilistic distribution of permitted cuts as part of the DNase-seq data, as well as extending the multivariate model so that it may account for sequence bias that is exhibited by the DNase I enzyme.

1. Introduction

1.1 Motivation

At any given time, thousands of genes may be in the process of being transcribed, and the coordination of such a massive undertaking relies on the specific pattern of regulatory elements binding to the cell's DNA. Since transcription is a temporally-dependent process, these regulatory pieces find themselves repeatedly changing positions along the genome. If we observe the realized configuration of these binding factors at several time points, we may glean an insight into the fundamental networks that drive the production of genes. Proper biological functioning depends heavily on the correctness and efficiency of these regulatory networks, and many diseases occur as a result of their misregulation, including cardiovascular disease, diabetes, and several types of cancer[1][2][3]. By developing models which integrate multiple sources of high-throughput sequencing data, we are able to more accurately predict the chromatin landscape of a given cell's genome.

Assaying techniques such as chromatin immunoprecipitation (CHiP) experiments have been successful in accurately locating the binding positions of regulatory proteins along the genome[4][5]. Unfortunately, conducting

the numerous CHiP experiments necessary to fully grasp the dynamics of the cell's chromatin would require an immense amount of resources, both in time and monetary cost, as a different experiment would be needed for each additional regulatory factor[6]. To this end, we propose a statistical inference model which accounts for potential biases in the newly integrated high-throughput DNase-seq data, with the goal of providing a more comprehensive overview of the chromatin state.

1.2 Related Work

The problem of discerning the binding locations of regulatory protein has been approached from multiple angles. Segal et al.[7] developed a thermodynamic model that predicts expression of cis-regulatory elements by using as input their respective binding profiles, along with the genome sequence. The model of transcription factor binding developed by Sinha[8] uses position weight matrices (PWMs) to discriminate between different binding locations for a given regulatory module. The COMPETE model[9] uses transcription factor binding specificity profiles along with the factors' concentrations to model the interaction of transcription factors and nucleosomes as they are competing for binding positions along the genome. In fact, it is the DNase-extended version of this model that we use as the starting point for modeling bias in the DNase-seq data, as we shall later see. An example of a multivariate hidden Markov model (MHMM) being used to understand the chromatin

profile, chromHMM[10] draws from multiple sources of chromatin marks data to make predictions about the chromatin state at 200-base-pair resolution.

2. Background

2.1 DNase I Enzyme

The DNase I enyzme is an endonuclease that cuts through the phosphodiester bond in DNA, thus leaving the DNA cleaved, and the position of this cleavage along the genome can be determined via sequencing methods[11]. In particular, DNase I binds to the minor groove of the DNA when making its cut[12], and it does so with a binding length of about 6-base-pairs[13]. Because the DNase I enzyme is more likely to cleave the DNA where the chromatin is unbound, knowledge of the frequency of cleavages reveals footprints of regulatory proteins, as detailed in the so-called DNase I-hypersensitivity data, also referred to as DNase-seq data[14]. In particular, Hesselberth et al.[14] propose a method for discerning the footprints of transcription factors, as well as their sequence motifs, that may be bound along the genome of *S. cerevisiae*. Their work exhibits the immense value that may be provided by the proper integration of DNase-seq data into a predictive statistical model, namely how the prevalence (or deficit) of cleaving can indicate that a dynamic binding factor resides (or is not found) at a certain position along the

genome.

2.2 Statistical Inference Methods

2.2.1 Hidden Markov Model (HMM)

A hidden Markov model (HMM) is a stochastic model whereby one can make predictions about the probability of realizing a finite set of unknown states (hence "hidden"), based on a finite sequence of observations that are emitted by the sequence of states that are traversed via the model, with exactly one emitted observation per traversed state[15]. More formally, an HMM is well-defined by its set of possible states $\pi = \{\pi_1, \pi_2, ..., \pi_m\}$, its set of possible observations $X = \{x_1, x_2, ..., x_n\}$, and the transitions and emissions probability distributions $\Pr_t$ and $\Pr_e$, respectively, where

$$\Pr_t = \Pr(\pi_j|\pi_i) \qquad \forall (i,j) \in \{1, ..., m\} \times \{1, ..., m\}$$

and

$$\Pr_e = \Pr(x_j|\pi_i) \qquad \forall (i,j) \in \{1, ..., m\} \times \{1, ..., n\}$$

Though perhaps a bit occluded by the mathematical formalism above, the inferential powers of HMMs cannot be overstated. One example is seen in how an HMM may be used to locate CpG islands along an unannotated sequence of nucleotides[16]. In COMPETE, as mentioned earlier, the states of the HMM represent either positions along a dynamic binding factor (including

nucleosomes and transcription factors), or the background state, all at single-nucleotide resolution, while the observations are the nucleotides along the genome sequence for the specified range and chromosome. It is important to note that each state in the original model emits one observation (i.e. there is exactly one observation node per state node), so with the introduction of an additional source of data to the model, each state will subsequently emit two observations.

2.2.2 Multivariate Extension to HMM

A natural extension to the HMM is to bring in additional sources of information as observations for the model, and in doing so, we end up with a multivariate hidden Markov model, or MHMM. Work currently being done in the group involves integrating DNase-seq hypersensitivity data as an additional observation in the COMPETE model, thus allowing for improved prediction of the chromatin occupancy profile as a result of better information about the potential for a given position along the genome being bound or unbound. A limitation of the current iteration, however, is that it assumes independence between the two observation sources, namely the genome sequence and the DNase-seq data. As we shall detail further later on, this independence assumption does not accurately reflect the nature of the data, as there is evidence that the emissions probability distribution from the hidden state to the DNase-seq observations depends on the surrounding sequence, with most

of the dependence reflected in a 6-base-pair window.

3. DNase I Sequence Bias

3.1 Observing DNase-seq Data

Before the integration of DNase-seq data into the model was possible, it was first necessary to determine the probability distribution of the emissions of the number of cleavages permitted by the chromatin for each position. Because the DNase-seq data is integral, its emission is modeled using a discrete probability distribution. Both Poisson and negative binomial distributions are especially well-suited to dealing with discrete count data, and so both options were considered as potential emissions distributions. Due to its only having a single parameter, the Poisson distribution fell short in its attempts to fit the data well. The negative binomial distribution, however, fared much better in describing the shape and distribution of the data, and thus it is the distribution used in the model.

In addition to using a negative binomial distribution to account for the emissions probabilities of the DNase-seq cleavage data, the current iteration of the model makes an independence assumption between the two observation data sources, thus implying that the amount of cleavage present at a given position does not depend on the genome sequence. Previous work has

shown, however, that this is a poor assumption to make, as the DNase I enzyme exhibits a strong sequence preference in its cleavage tendencies[17][18]. After analyzing the in vitro DNase I-hypersensitivity data of the *S. cerevisiae* genome from Hesselberth et al.[14], we observed a similarly meaningful sequence bias (Figure 3.1). Interestingly, a majority of the sequence specificity is contained within approximately a 6-base-pair region around the cleavage site, an artifact we would expect since it is the binding length of the DNase I enzyme[13].

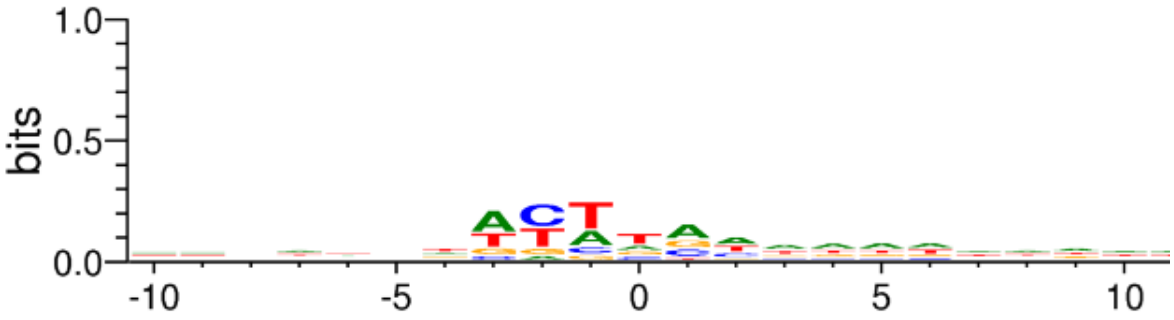

Figure 3.1: Position weight matrix (PWM) displaying the sequence bias. The cleavage made by DNase I enzyme is present between positions -1 and 0 in the diagram.

3.2 Accounting for Bias in Model

Now that we have established a clear sequence bias exists in the cleavage patterns of the DNase I enzyme, the logical next step is to make changes in the model which account for this new information. In a sense, we need to adjust

the topology of the multivariate HMM so that the DNase-seq observed count data depends on both the underlying state and the local genome sequence, thus breaking the independence assumption (Figure 3.2). Our chosen method for incorporating the bias knowledge into the model begins by observing a small segment of the genome surrounding the cleavage position. Because of the DNase I enzyme's binding length, as well as the patterns observed before from our analysis of the in vitro data, we use a 6-mer model of the local sequence.

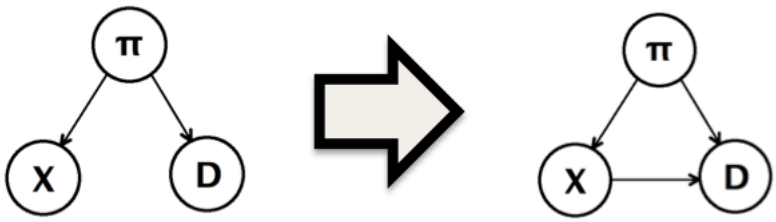

Figure 3.2: Desired transformation in adjusting the multivariate hidden Markov model topology, and removing the independence assumption from the model. π is the chromatin state sequence, X is the genome sequence, and D is the set of DNase-seq cleavage amounts.

To keep track of whether a certain motif is over- or under-cut, we maintain a mapping from every possible 6-mer to a ratio of its experimentally determined cleavage amount, based on the in vitro data from Hesselberth et al.[14], over its expected cleavage amount, and we call this value the bias term, denoted b, for every possible 6-mer. Tables 3.1 and 3.2 list the most over-selected and under-selected 6-mers, respectively, along with their bias

6-mer	Observed Cuts / Expected Cuts
TCTTAC	24.211
ACTTAA	22.364
ACTTAC	21.636
ACTTAT	19.928
ACTTGA	18.136
TCTTAT	17.282
ACTTGT	16.150
TCTTAA	14.748
ACTTGC	14.166
GCTTAC	13.498

Table 3.1: Top 10 Over-Biased 6-mers

terms. In addition, Figure 3.3 exhibits the distribution of the log of the bias terms for all of the 6-mers.

6-mer	Observed Cuts / Expected Cuts
CACCGG	0.0119
GACCGG	0.0134
CGCGTG	0.0142
CGCCGG	0.0152
CACATG	0.0175
CAGGCG	0.0176
CACTTT	0.0181
CACGCG	0.0206
GACTTT	0.0212
GACATT	0.0218

Table 3.2: Top 10 Under-Biased 6-mers

Given the surrounding 6-mer for each position along the genome, we adjust the emissions distribution for the DNase-seq data, regardless of the original parameters of its negative binomial distribution (which depend solely on the hidden occupancy state), as follows:

Suppose we want to determine the probability distribution of the chromatin landscape for a sequence of length N. Then, for $i = (4, ..., N-2)$, indexing by 1, we determine our local 6-mer nucleotide window around position i, denoted x_{local}, from the nucleotides in the range $(i-3, ..., i+2)$, inclusive. Let

$$NB(n_i, p_i)$$

represent the negative binomial emissions probability distribution from state π_i to the DNase-seq cleavage count at position i, denoted d_i. Thus,

$$\Pr(d_i \; cuts \; emitted) = \binom{d_i + n_i - 1}{d_i} \cdot (1-p_i)^{n_i} p_i^{d_i}$$

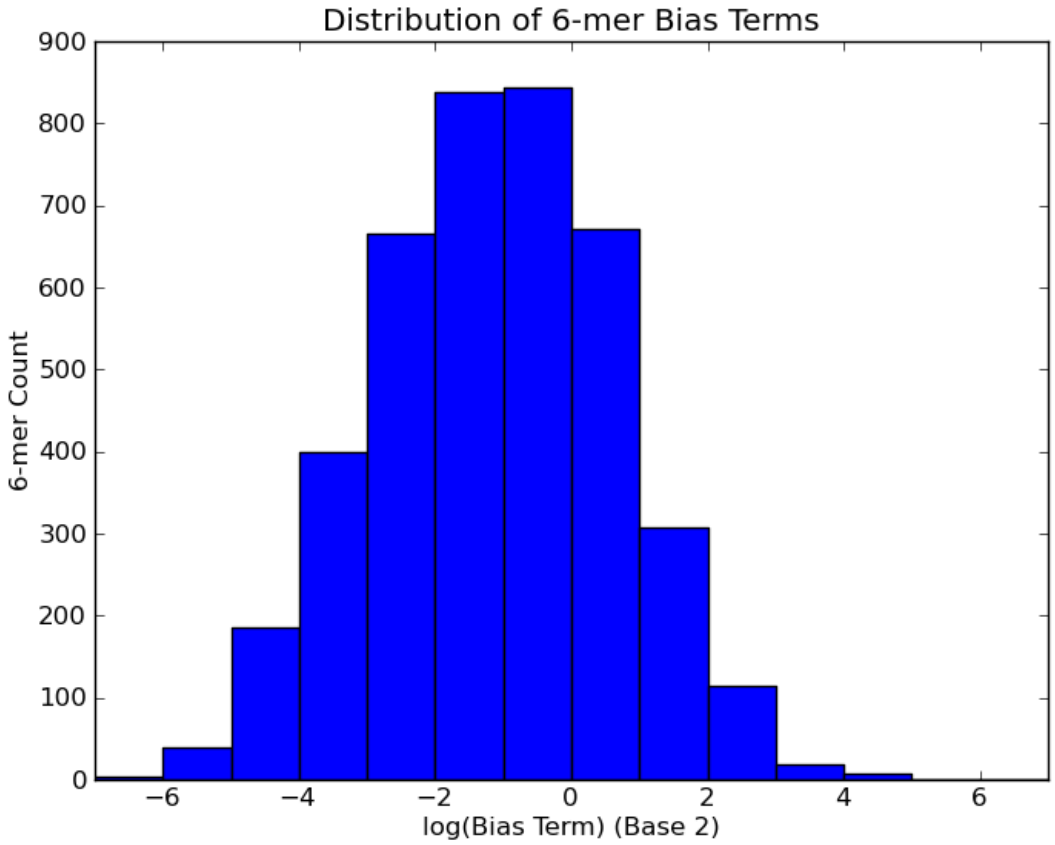

Figure 3.3: Distribution of (the log of) the bias terms, as determined by the in vitro DNase-seq data of the *S. cerevisiae* genome from Hesselberth et al.[14].

The values of n_i and p_i are assumed known, based on the original version of the multivariate HMM, and they are also sufficient to determine the mean of the distribution, μ_i, since

$$\mu_i = \frac{n_i(1-p_i)}{p_i} \, .$$

The current formulation is such that

$$\Pr(d_i, x_i|\pi_i) = \Pr(d_i|\pi_i)\Pr(x_i|\pi_i)$$

Now, given the surrounding nucleotide window for position i, denoted x_{local_i}, we may use our lookup table generated before to find the corresponding bias term, b_i. Our goal is to scale the shape of the distribution, and we accomplish this by applying a scaling function

$$s(x, j) = x^j$$

for a chosen j, $0 \leq j \leq 1$, to our current bias term b_i. Doing so allows us to "tune" the weight of the bias term (whereby $j = 1$ implies no change to the bias term, and $j = 0$ changes all of the bias terms to 1) while maintaining a factor which is consistent with the bias term's purpose of shifting the negative binomial distribution to either over- or under-predict the number of DNase-seq cuts. In this version of the model, we found that choosing $j = 0.1$ allowed for the scaling factor to have a meaningful effect, without "overpowering" the resulting distribution. Thus, we create a modified bias term,

$$\widehat{b_i} = s(b_i, 0.1)$$

and use this term to determine the new parameters of our adjusted negative binomial distribution:

$$\widehat{\mu_i} = \mu_i \widehat{b_i}$$

$$\widehat{p_i} = \frac{n_i}{n_i + \widehat{\mu_i}}$$

Finally, our new DNase-seq cuts emissions distribution is

$$\widehat{\Pr}(d_i|\pi_i) \sim NB(n_i, \widehat{p_i})$$

and with this in hand, we make the following modifications for our model:

$$\Pr(d_i, x_i|\pi_i) = \widehat{\Pr}(d_i|\pi_i) \Pr(x_i|\pi_i)$$

In taking these steps, we are able to account for the bias relating to any specified 6-mer, without introducing a series of actual additional edges in the model. Instead, since the genome sequence information is fully known beforehand, we may generate the scaled distribution without any knowledge of the hidden states, probability distributions, etc.

By scaling the parameters of the negative binomial distributions, as described here, we allow for under- or over-biased positions to be compensated accordingly. In other words, under-biased positions will result in a negative binomial distribution with a smaller mean for the number of DNase-seq cuts, thus creating a probability distribution that more accurately reflects

the number of cuts to be expected from that position, while the opposite holds true for over-biased positions.

4. Results

After incorporating this modification into the model in order to account for sequence bias, we may look at how the chromatin occupancy prediction of the bias-including model compare to the original, bias-ignoring, multivariate model. Upon observing a specific locus of the *S. cerevisiae* genome, namely positions 636k to 638k on chromosome XIII (Figure 4.1), we are immediately witness to the differences in the binding factor predictions of the two versions of the model. These differences are more pronounced in the case of the nucleosomes, as both models correctly predict the positions of two Reb1 transcription factors at known Reb1 binding sites[4].

For the bias-including model, we notice the placement of an additional nucleosome which is mostly absent from the bias-ignoring model, and there is a signal, albeit weak, from the Brogaard et al. data predicting a nucleosome center at that position. Also clearly visible for the bias-including model is an additional smoothing of the nucleosome prediction probabilities, when compared to the rougher terrain along the nucleosomes as demonstrated by the bias-ignoring plot. Finally, the strongest signal from the Brogaard et al. data in the region, just after position 637k, is more confidently predicted by

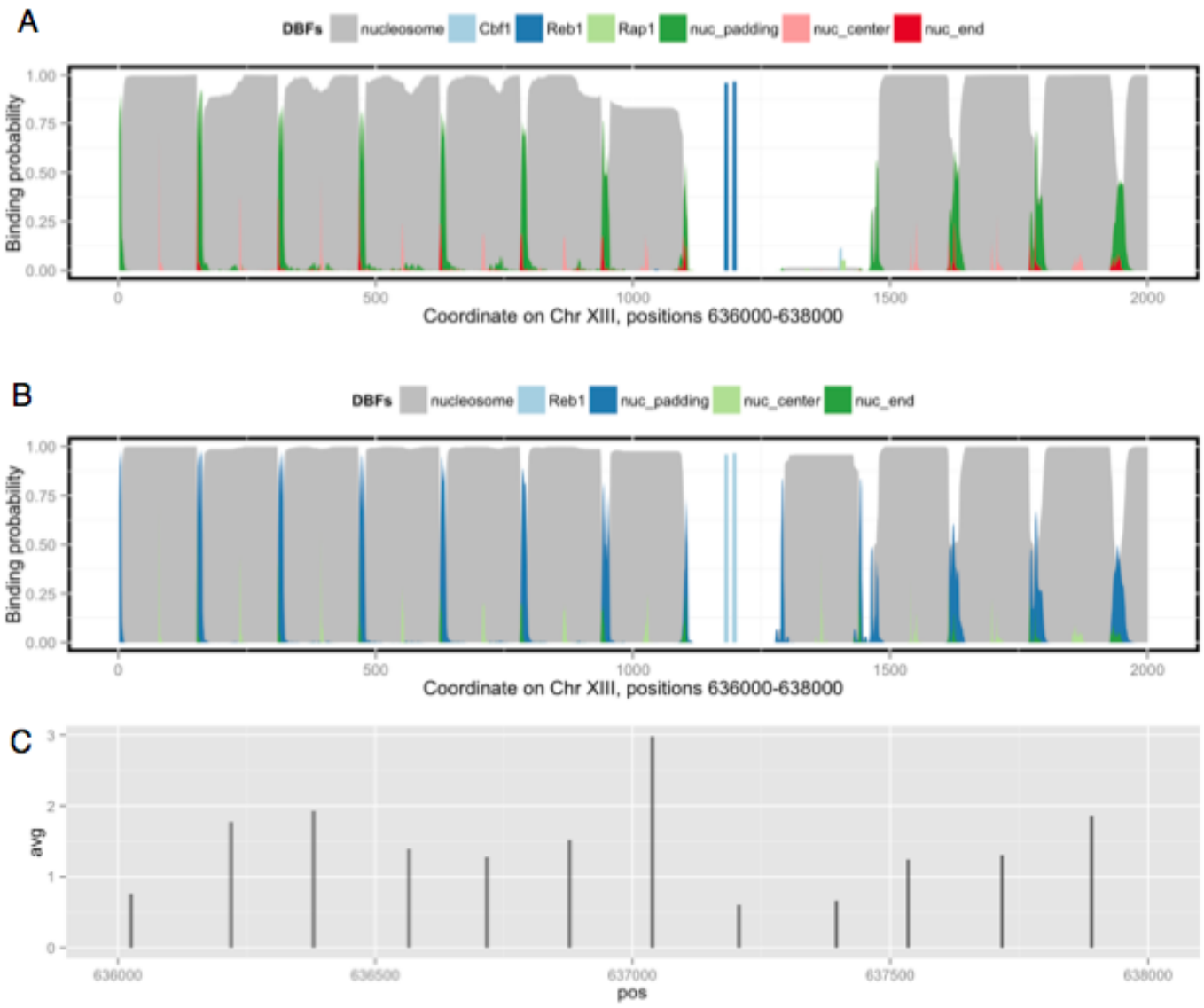

Figure 4.1: Comparison of chromatin binding profile predictions to predicted nucleosome positions. (A) Profile prediction, bias-ignoring model. (B) Profile prediction, bias-including model. (C) Confidence of nucleosome center positions, from the NCP score-to-noise ratio from Brogaard et al.[19].

Model Version	Average Nucleosome Predicted Probability
Bias Not Included	0.90735285
Bias Included	0.91638405

Table 4.1: Genome-wide Analysis Comparing Predicted and Experimental Nucleosome Positioning

the bias-including model than the bias-ignoring model.

To test the accuracy of our model's predictions genome-wide, we considered all 349 nucleosome centers in the unique map from Brogaard et al.[19] with an NCP score-to-noise ratio of ≥ 10. We then focused on a 100-base-pair window, 50-base-pairs on either side, around each of the 349 nucleosome centers, and averaged the predicted binding probabilities of the nucleosome binding factor for all 100-base-pair windows over all specified nucleosome centers, for both the bias-ignoring and the bias-including models. After running this analysis, we obtained the results as shown in Table 4.1.

We may note that the bias-including model fared slightly better in predicting nucleosome occupancy around highly-probable experimentally determined nucleosome center positions, thus providing another indication of the efficacy of including a modification in the model to account for the sequence

specificity of the DNase I enzyme.

5. Future Work

The current model, in accounting for bias, has helped bring the appropriate dependence relationship between the local sequence segment and the DNase-seq cleavage data. One possibility for improvement would be to include more sophisticated methods to deconvolve the bias factor for a given setting of the model. I chose against this idea to avoid overcomplexity, but perhaps there is a configuration which may include additional (true) edges in the multivariate HMM that bridge connections between individual observation nodes. Of course, in order for this to work, it would need to be accomplished without overcomplicating the application of standard inference algorithms to the model, while still producing reliable predictions of the chromatin occupancy profile.

Another direction in which the model could be taken would be to integrate additional data sets that may lend a hand to predicting chromatin occupancy. A promising source of data that may provide insight into the chromatin state is generated by yet another enzyme, MNase. The data sets generated as a result of MNase-digested DNA are especially of interest as they detail entire fragment of DNA which remained under enough protection (ostensibly due to the presence of binding factors) to forgo any further digestion, thus providing

information about the position and length of protected regions along the genome[20].

6. Conclusion

Due to the importance of a cell's chromatin state with respect to the proper regulation of its genes' transcription, it would prove useful to have a system which can model the binding patterns of regulatory proteins. One such model, a multivariate extension of COMPETE, currently uses as input both the genome sequence and DNase-seq cleavage data, whereby the output is a profile describing the binding probabilities of different factors at single-nucleotide resolution. A key independence assumption made by this version of the model concerning the genome sequence and the DNase-seq data set, however, begins to falter when presented with evidence among other data sets which indicates sequence bias of the DNase I enzyme, and ultimately crumbles upon observation of the bias that is exhibited within the in vitro DNase-seq data set from the *S. cerevisiae* genome.

Our modified model accounts for this bias by carefully adjusting the parameters of the emissions distribution based on a small 6-base-pair segment of the genome sequence surrounding each position. After correcting for bias, we observe results in improved prediction of binding probabilities for dynamic binding factors, especially in the case of predicting the positions of

nucleosomes. These results appear both at specific loci where the dynamic binding profiles are compared, as well as through comprehensive genome-wide comparison of the average nucleosomal prediction at locations with high likelihood of nucleosome center positioning.

References

[1] T.I. Lee and R. A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152:1237-1251, 2013.

[2] Y. Jiang, H. Shen, X. Liu, J. Dai, G. Jin, et al. (2011). Genetic variants at 1p11.2 and breast cancer risk: a two-stage study in Chinese women. *PLoS ONE* 6:e21563, 2011.

[3] F. W. Huang, E. Hodis, M. J. Xu, G. V. Kryukov, L. Chin, and L. A. Garraway. Highly recurrent TERT promoter mutations in human melanoma. *Science* 339:957-959, 2013.

[4] H. S. Rhee and B. F. Pugh. Comprehensive genome-wide protein-DNA Interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408-1419, 2011.

[5] A. Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*, 16:962-972, 2006.

[6] M. J. Solomon and A. Varshavsky. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proceedings of the National Academy of Sciences*, 82:6470-6474, 1985.

[7] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451:535-540, 2008.

[8] S. Sinha. On counting position weight matrix matches in a sequence, with application to discriminative motif nding. *Bioinformatics*, 22(14):e454-e463, 2006.

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

[10] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*, 9:215-216, 2012.

[11] A. Lazarovici, T. Zhou, A. Shafer, A. C. D. Machado, T. R. Riley, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proceedings of the National Academy of Sciences*, 110(16):6376-6381, Apr 2013.

[12] D. Suck, A. Lahm, and C. Oefner. Structure refined to 2A of a nicked DNA octanucleotide complex with DNase I. *Nature*, 332(6163):464-468, 1988.

[13] S. A. Weston, A. Lahm, and D. Suck. X-ray structure of the DNase I-d(GGTATACC)2 complex at 2.3 A resolution. *J Mol Biol*, 226(4):1237-1256, 1992.

[14] J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, 6(4):283-289, Apr 2009.

[15] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257-286, Feb 1989.

[16] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Cambridge: Cambridge University Press, 1998. Print.

[17] H. H. He, C. A. Meyer, S. S. Hu, M.-W. Chen, C. Zang, et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods*, 11(1):73-78, Jan 2014.

[18] H. Koohy, T. A. Down, and T. J. Hubbard. Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. *PLoS ONE* 8(7):e69853, 2013.

[19] K. Brogaard, L. Xi, J.-P. Wang, and J. Widom. Map of Nucleosome Positions in Yeast at base-pair resolution. *Nature*, 486:496-501, 2012.

[20] J. G. Henikoff, J. A. Belsky, K. Krassovsky, D. M. MacAlpine, and S. Henikoff. Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences*, 108(45):1831818323, 2011.