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

DUKE COMPUTER SCIENCE

Brian Bullins and Alexander Hartemink
Duke University, Department of Computer Science

Introduction

•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.

•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 (Lee & Young, 2013).

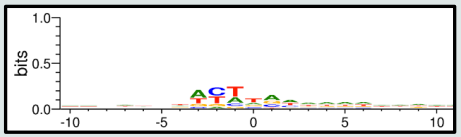
•We propose an extended statistical inference model which accounts for potential biases in the newly integrated high-throughput DNase-seq data, with the goal of providing a more comprehensive overview of the chromatin state.

DNase I Enzyme Bias

•DNase I is an endonuclease that cleaves DNA along the minor groove, with a window of about 6 recognized base pairs (Lazarovici et al., 2013).

•It has been shown for other data sets that DNase I exhibits sequence specific bias in where it tends to cleave the DNA (Koohy et al., 2013).

•After analyzing the in vitro DNase-seq hypersensitivity data, we observed a similar instance of bias for the enzyme, as displayed by the PWM below.



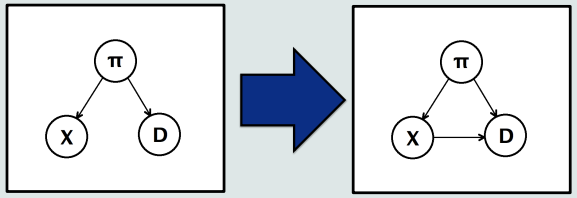
DNase I cleavage is between positions -1, 0

Extending the Multivariate Model

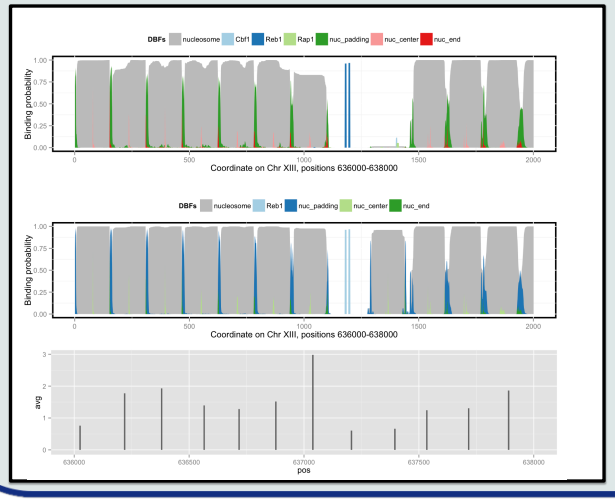
•Multivariate version of COMPETE uses both genome sequence data, X, and DNase-seq cleavage data, D, as observation nodes in the hidden Markov model.

•Current iteration assumes independence between the observation nodes, but given the presence of sequence bias in the DNase-seq data, this is a poor assumption. Thus, we need to somehow account for this bias within the model, as demonstrated by the diagram below.

•Solution: Given a relevant window of the genome sequence surrounding the cleavage position (our model uses a 6-mer window), adjust the emissions probability distribution of the DNase-seq data for each individual position.



Resulting Improvements in Predicting Chromatin State



•Focusing on a specific locus, we may observe the advantage gained in the model's chromatin occupancy predictive capabilities.

•The top two plots represent the binding probabilities for different binding factors along chromosome XIII from positions 636k to 638k—the topmost plot represents the model not accounting for bias, and the middle plot represents the model accounting for bias.

•When accounting for bias, we may notice the placement of an additional nucleosome which is absent in the first plot.

•We may also notice the "smoothing" of nucleosomes on the left-hand side of the bias-including plot, compared to the bias-ignoring plot.

•To gauge the accuracy of the model's predictions, we have included a plot of the positions of experimental nucleosome centers along the specified region (Brogaard et al., 2012).

•The most confidently placed nucleosome, according to the Brogaard data, is better predicted as such by the bias-including model than by the bias-ignoring model.

Conclusion

•When integrating DNase-seq hypersensitivity data as part of the multivariate version of COMPETE, it is important to consider sequence specific bias exhibited by the Dnase I enzyme.

•Our model accounts for this bias by carefully adjusting the parameters of the emissions distribution, based on the information of the local sequence data.

•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.

References

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

Koohy, H., 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.

Lazarovici, A., 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. *PNAS*, 110(16):6376-6381, Apr 2013.

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

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