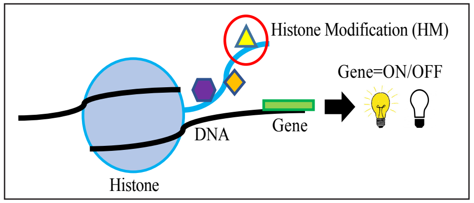
Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin

Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, & Yanjun Qi
University of Virginia, Department of Computer Science

1. Overview

Gene regulation is the process of how the cell controls which genes are turned "on" (expressed) or "off" (not-expressed) in its genome. ``Chromatin" denotes DNA and its organizing proteins. A cell uses specialized proteins to organize DNA in a condensed structure. These proteins include histones, which form "bead"-like structures that DNA wraps around, organizing and compressing the DNA. An important aspect of histone proteins is that they are prone to chemical modifications that can change the spatial arrangement of DNA. These spatial re-arrangements result in certain DNA regions becoming accessible or restricted and therefore affecting expressions of genes in the neighborhood region.



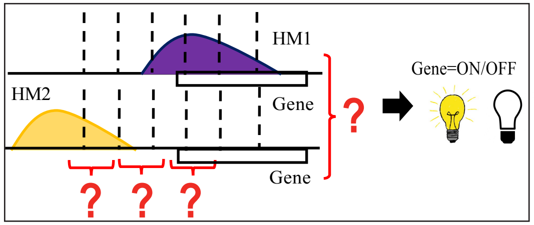
Researchers have established the "Histone Code Hypothesis" that explores the role of histone modifications (HMs) in controlling gene regulation.

2. Challenges

Recent literature tried to understand gene regulation by predicting gene expression from large-scale chromatin measurements. Two fundamental challenges exist for such learning tasks:

(1) Genome-wide chromatin signals are spatially structured, high-dimensional and highly modular

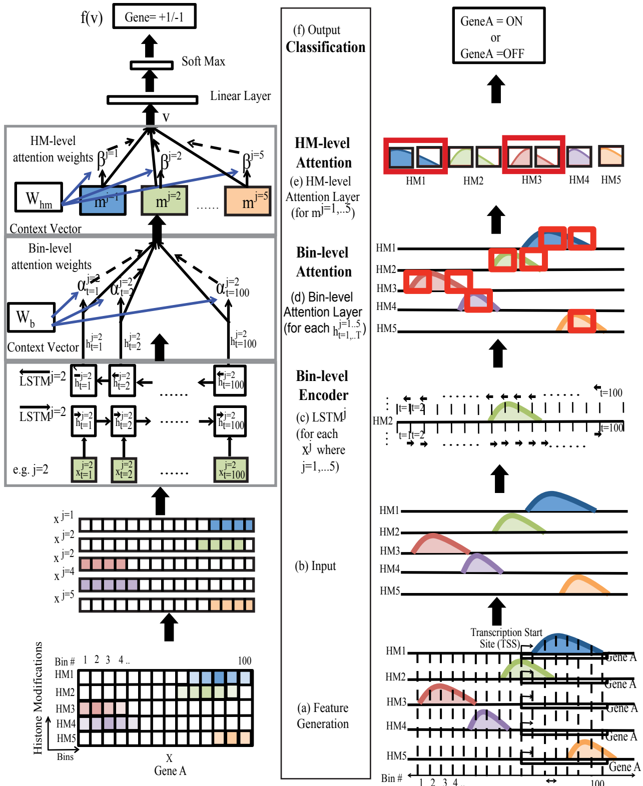
(2) The core aim is to understand what the relevant factors are and how they work together.



Previous studies either failed to model complex dependencies among input signals or relied on separate feature analysis to explain the decisions.

3. Approach

We present an attention-based deep learning model, AttentiveChrome, that uses a hierarchy of multiple Long Short-Term Memory (LSTM) modules [1] to encode the input signals and to model how various chromatin marks cooperate automatically. AttentiveChrome trains two levels of attention jointly with the target prediction, enabling it to attend differentially to relevant marks and to locate important positions per mark.



References:
[1] Zichao Yang et al. Hierarchical attention networks for document classification. HLT. 2016
[2] Anshul Kundaje et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015
[3] Ritambhara Singh et al. DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinfromatics. 2016.

Code Available at: https://github.com/QData/AttentiveChrome

Contact: rs3zz@virginia.edu

Acknowledgement: We would like to thank the NIPS foundation for the Travel Award..

4. Experiments and Results

We downloaded gene expression levels and signal data of five core HM marks for 56 different cell types archived by the REMC database [2]. Each dataset contains information about both the location and the signal intensity for a mark measured across the whole genome.

These five HM marks include (we rename these HMs in our analysis for readability):

Histone Modification (HM) Mark	Renamed as	Functional Category
H3K4me3	H_{prom}	Promoter mark
H3K4me1	H_{enhc}	Distal Enhancer mark
H3K36me3	H_{struct}	Structural mark
H3K9me3	H_{reprA}	Repressor mark
H3K27me3	H_{reprB}	Repressor mark

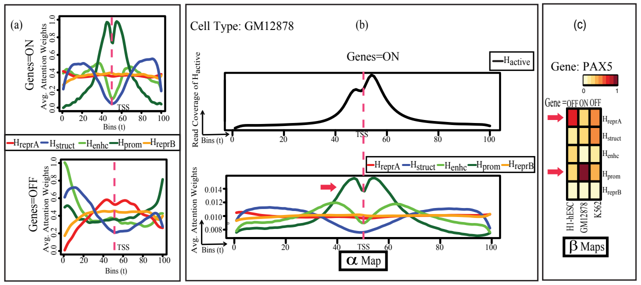
We compare AttentiveChrome using summarized AUC scores across all 56 cell types on the test set. We find that overall the Attentive-Chrome performs better than CNN-based [3] and LSTM baselines.

	Baselines		Our Model
Models	**DeepChrome (CNN) [3]**	**LSTM**	**AttentiveChrome**
Mean	0.8008	0.8052	**0.8115**
Median	0.8009	0.8036	**0.8123**
Max	**0.9225**	0.9185	0.9177
Min	0.6854	0.7073	**0.7215**
Improvement over DeepChrone [3] (out of 56 cell types)		36	**49**

Next, we demostrate that AttentiveChrome allows interpretability to the "black box" neural networks.

(a) Bin-level attention weights α from AttentiveChrome averaged for all genes when predicting gene=ON and gene=OFF in cell-type GM12878 (blood cell).

(b) Unlike images and text, the results for biology are hard to interpret by just looking at them. We use additiona signal - H3K27ac (H_{active}) from REMC database [2]. This HM marks the region that is active when the gene is "ON". We show cumulative H_{active} signal across all active genes. H_{prom} α weights for gene=ON correspond well with the H_{active} indicating actual activity near the gene. This shows that AttentiveChrome is focusing on the correct bin positions for this case



(c) Heatmaps visualizing the HM-level weights β, for an important differentially regulated gene (PAX5) across three blood lineage cell types: H1-hESC (stem cell), GM12878 (blood cell), and K562 (leukemia cell). The trend of HM-level β weights for PAX5 have been verified through biological literature.