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 rearrangements 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.

3. Approach
We present an attention-based deep learning model, AttentiveChrome, that uses a hierarchy of multiple Long Short-Term Memory (LSTM) modules 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.

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):

- Gene A
- Gene B
- Gene C
- Gene D
- Gene E

Next, we demonstrate that AttentiveChrome allows interpretability to the “black box” neural networks.

(a) Bin-level attention weights $\alpha$ from AttentiveChrome averaged for all genes when focusing on the correct bin positions for this case

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

References:

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

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