Making **Deep Learning Understandable** for Analyzing **Sequential Data** about **Gene Regulation**

Dr. Yanjun Qi

2017/11/26
Roadmap

✧ Background of Machine Learning
✧ Background of Sequential Data about Gene Regulation
✧ AttentiveChrome for understanding gene regulation by selective attention on chromatin
Roadmap

✧ Background of Machine Learning
✧ Background of Sequential Data about Gene Regulation
✧ AttentiveChrome for understanding gene regulation by selective attention on chromatin
Machine Learning is Changing the World

How may I help you, human?

Apple Siri / Amazon Echo

Many more!

Object recognition

Control learning
Challenge of data explosion

Molecular signatures of tumor / blood sample

- Signs & Symptoms
- Patient Medical History & Demographics
- Public Health Data
- Genetic Data
- Mobile medical sensor data
- Medical Images

Traditional Approaches

Data-Driven Approaches

Machine Learning
What is Machine Learning?

Input: \( X \)

Output: \( Y \)

Function: \( f(X) \)
What is Machine Learning?

Input: $X'$  \[\rightarrow\]  \text{Function} $f(X)$  \[\rightarrow\]  Output: $Y'$
[Example:] What is Machine Learning?

Output: CAR
Inputs:

Training

Function $f(X)$
[Example:] What is Machine Learning?
[Example:] Classification task in Machine Learning

Class: Car

Testing

Function $f(X)$

NO (-1)
Sequential Data

Strings, signals etc.

This food is not good.
This Food is **not good**.

[Example:] Classification of Sequential Data

\[ x \rightarrow f(X) \rightarrow \text{NO (-1)} \]
State-of-the-art Machine Learning - Deep Neural Networks (DNN)
State-of-the-art Machine Learning - Deep Neural Networks (DNN)

“Dog”

Can get overly sentimental at times, but Gus Van Sant’s sensitive direction... and his excellent use of the city make it a hugely entertaining and effective film.

Full Review... | May 25, 2006
State-of-the-art Machine Learning - Deep Neural Networks (DNN)

Can get overly sentimental at times, but Gus Van Sant's sensitive direction... and his excellent use of the city make it a hugely entertaining and effective film.

Full Review... [May 25, 2006]

ATGC\textcolor{red}{G}AT\textcolor{blue}{C}AAGTCTG

“Protein Binding Site”
Deep Neural Networks (DNN)

\[ Y = f_4 (f_3 (f_2 (f_1 (X)))) \]
ATGCGATCAAGTCTG → “Protein Binding Site” → “Dog” → Can get overly sentimental at times, but Gus Van Sant’s sensitive direction... and his excellent use of the city make it a hugely entertaining and effective film.

Full Review... | May 25, 2006
DNNs are hard to Interpret

Our Goals

Accurate

Understandable
Roadmap

✧ Background of Machine Learning
✧ Background of Sequential Data about Gene Regulation
✧ AttentiveChrome for understanding gene regulation by selective attention on chromatin
Biology in a Slide

Transcription

DNA

RNA

PROTEIN

Translation

CELL

ORGANISM
Genome Organization and Gene Regulation

- Regulatory Elements
  - Genes
  - Promoters
  - Enhancers

- Chromatin Structure
  - Histone Modifications
  - DNA methylation
  - Chromatin remodeling

- Nuclear Architecture
  - Chromosomal organization
  - Long range interactions

Cellular Phenotype

(adapted from Babu et al., 2008)
ENCODE Project (2003-Present)
Describe the functional elements encoded in human DNA
ENCODE Project (2003-)

Describe the functional elements encoded in human DNA

Roadmap Epigenetics Project (REMC, 2008-)

To produce a public resource of epigenomic maps for stem cells and primary ex vivo tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease.
Current Available Large-Scale Data about Gene Transcription

DNA Segments on Genomes

TF Binding Signals

Histone Modification Signals

Gene Expression

ATCGATCAAGTCTG
Two Important Data-Driven Computational Tasks

DNA Segments on Genomes

TF Binding Signals

Histone Modification Signals

Gene Expression

This Talk

ATGCCGATCAAGTCTG
One main aim of such data analysis is to understand data and to discover knowledge.
Roadmap

✧ Background of Machine Learning
✧ Background of Sequential Data about Gene Regulation
✧ AttentiveChrome for understanding gene regulation by selective attention on chromatin
Chromatin

DNA
histone
TFs

Control

Gene Expression
Transcription Factor Binding => Gene Transcription
Histone Modifications (HM)
Histone Modification and Gene Transcription

Transcription Factor (TF)

↑

Histone Modification (HM)

Gene Transcription
Histone Modification and Gene Transcription

- Transcription Factor (TF)
- Histone Modification (HM)
- Gene Transcription

?
Histone Modification and Gene Transcription

Transcription Factor (TF)

Histone Modification (HM)

Gene Transcription

?
Why Studying [HM => Gene Expression]?

• Epigenomics:
  • Study of chemical changes in DNA and histones (without altering DNA sequence)
  • Inheritable and involved in regulating gene expression, development, tissue differentiation and suppression …

• Modification in DNA/histones (changes in chromatin structure and function)
  • => influence how easily DNA can be accessed by TF

• Epigenome is dynamic
  • Can be altered by environmental conditions
  • Unlike genetic mutations, chromatin changes such as histone modifications are potentially reversible => Epigenome Drug for Cancer Cells?
Study what HMs affect which genes in what cells?
Task Formulation

Prediction Task

Input:

HM1

HM2

HM3

HM4

HM5

Histone Modification Signals

Output: Gene ON/OFF
Histone Modification Signals

HM1

HM2

HM3

HM4

HM5

DNA

Gene

Transcription Start Site
Challenge
Challenge

Histone Modification Signals

HM1
HM2
HM3
HM4
HM5

Transcription Start Site

Gene

DNA
Computational Challenge

Search Space per Gene = $2^{100 \times 5}$
Related Work

Linear Regression, SVM, Random Forest

Gene ON/OFF
Drawback of Related Works

Gene ON/OFF

Linear Regression, SVM, Random Forest
Deep Learning to Rescue:

ImageNet Error Rate

Deep Neural Network (DNN)

Using deep learning

Human performance
Deep Learning to Rescue:

Input

- SKY
- TREE
- HUMAN

Output

- Park

Gene

HM1
DNA

HM5
DNA
DeepChrome-Convolutional Neural Network (CNN)

1. Convolution
2. Max Pooling
3. Dropout
4. Multi-Layer Perceptron
5. Softmax

\[ L = \sum_{n=1}^{N_{\text{samp}}} \text{loss}(f(X^{(n)}), y^{(n)}) \]
DeepChrome-Convolutional Neural Network (CNN)

Back-propagation:
\[ \Theta \leftarrow \Theta - \eta \frac{\partial L}{\partial \Theta} \]
Experimental Setup

- **Cell-types:** 56
- **Input (HM):** ChIP-Seq Maps (REMC)
- **Output (Gene Expression):** Discretized RNA-Seq (REMC)

<table>
<thead>
<tr>
<th>Histone Mark</th>
<th>Functional Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K27me3</td>
<td>Repressor</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>Promoter</td>
</tr>
<tr>
<td>H3K4me1</td>
<td>Distal Promoter</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Promoter</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>Repressor</td>
</tr>
</tbody>
</table>

- **Baselines:** Support Vector Classifier (SVC) and Random Forest Classifier (RFC)

Training Set 6601 Genes  Validation Set 6601 Genes  Test Set 6600 Genes
Results: Accuracy

- First deep learning implementation for gene expression prediction.
- But hard to interpret.
Accurate

DeepChrome

Understandable

AttentiveChrome: understanding gene regulation by selective attention on chromatin

Goal: one DNN both accurate and interpretable

Interpretability by Attention

Input

Output

HM1
DNA
Gene

HM2
DNA
Gene

Attention Mechanism

Park

Gene

54
Two Levels of Attention

Bin-level Attention

HM1 ($X^j=1$)
HM2 ($X^j=2$)
HM3 ($X^j=3$)
HM4 ($X^j=4$)
HM5 ($X^j=5$)

Transcription Start Site (TSS)

Gene A

Bin #

$\alpha^j_t$

$\beta^j$

HM-level Attention

Gene A = ON/OFF ($y=+1/-1$)

Classification
Overview

Input

HM1

HM2

HM3

LOCAL LEVEL
Overview

GLOBAL LEVEL

LOCAL LEVEL

Input

HM1

HM2

HM3
Overview

GLOBAL LEVEL

Input

HM1

LOCAL LEVEL

Gene

Output

61

HM2

HM5
Recurrent Neural Network (RNN)

HM1
Recurrent Neural Network (RNN)

Attention Mechanism

HM1
## Versus Baselines

<table>
<thead>
<tr>
<th>Computational Study</th>
<th>Unified</th>
<th>Non-linear</th>
<th>Bin-Info</th>
<th>Representation Learning</th>
<th>Prediction</th>
<th>Feature Inter.</th>
<th>Interpretable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neighbor Bins</td>
<td>Whole Region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Regression ([14])</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Support Vector Machine ([7])</td>
<td>×</td>
<td>✓</td>
<td></td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Random Forest ([10])</td>
<td>×</td>
<td>✓</td>
<td>Bin-specific</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Rule Learning ([12])</td>
<td>×</td>
<td>✓</td>
<td>Best-bin</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>DeepChrome-CNN [29]</td>
<td>✓</td>
<td>✓</td>
<td>Automatic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>AttentiveChrome</strong></td>
<td>✓</td>
<td>✓</td>
<td>Automatic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
### Experiments: Prediction Performance

- Same setup as DeepChrome
- AttentiveChrome is as accurate as (slightly better than) DeepChrome

<table>
<thead>
<tr>
<th>Models</th>
<th>Baselines</th>
<th>Our Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>DeepChrome (CNN) [3] 0.8008</td>
<td>LSTM 0.8052</td>
</tr>
<tr>
<td>Median</td>
<td>DeepChrome (CNN) [3] 0.8009</td>
<td>LSTM 0.8036</td>
</tr>
<tr>
<td>Max</td>
<td>DeepChrome (CNN) [3] 0.9225</td>
<td>LSTM 0.9185</td>
</tr>
<tr>
<td>Min</td>
<td>DeepChrome (CNN) [3] 0.6854</td>
<td>LSTM 0.7073</td>
</tr>
<tr>
<td>Improvement over DeepChrome [3] (out of 56 cell types)</td>
<td>36</td>
<td>49</td>
</tr>
</tbody>
</table>
Experiments: Interpretability

- Local-level (HM-level) Attention
- Global-level (HM interactions) Attention

<table>
<thead>
<tr>
<th>Histone Modification (HM) Mark</th>
<th>Renamed as</th>
<th>Functional Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me3</td>
<td>H\text{prom}</td>
<td>Promoter mark</td>
</tr>
<tr>
<td>H3K4me1</td>
<td>H\text{enhc}</td>
<td>Distal Enhancer mark</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>H\text{struct}</td>
<td>Structural mark</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>H\text{reprA}</td>
<td>Repressor mark</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>H\text{reprB}</td>
<td>Repressor mark</td>
</tr>
</tbody>
</table>
(1) Visualization of Local Attention Weights (Learned from Data)

- Additional signal - H3K27ac (H-Active) from REMC
- Average local attention weights of gene=ON correspond well with H-active
- Indicating AttentiveChrome is focusing on the correct bin positions
An important differentially regulated gene (PAX5) across three blood lineage cell types:
- H1-hESC (stem cell),
- GM12878 (blood cell),
- K562 (leukemia cell).

Trend of its global weights (beta) Verified through the literature.
(3) Comparison with State-of-Art Deep-Visualization Methods

Correlation between local-level (HM-level) attention weights and the additional signal - H3K27ac (H-Active) from REMC

Table 3: Pearson Correlation values between weights assigned for $H_{prom}$ (active HM) by different visualization techniques and $H_{active}$ read coverage (indicating actual activity near "ON" genes) for predicted "ON" genes across three major cell types.

<table>
<thead>
<tr>
<th>Viz. Methods</th>
<th>H1-hESC</th>
<th>GM12878</th>
<th>K562</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$ Map (LSTM-$\alpha$)</td>
<td>0.8523</td>
<td>0.8827</td>
<td>0.9147</td>
</tr>
<tr>
<td>$\alpha$ Map (LSTM-$\alpha$, $\beta$)</td>
<td><strong>0.8995</strong></td>
<td>0.8456</td>
<td>0.9027</td>
</tr>
<tr>
<td>Class-based Optimization (CNN)</td>
<td>0.0562</td>
<td>0.1741</td>
<td>0.1116</td>
</tr>
<tr>
<td>Saliency Map (CNN)</td>
<td>0.1822</td>
<td>-0.1421</td>
<td>0.2238</td>
</tr>
</tbody>
</table>
Summary

- **Attentive DeepChrome**
  - Both accurate and interpretable
  - Novel implementation of deep attention mechanism
  - Importance analysis at both HM and HM-HM level

code available at: deepchrome.org
References


Acknowledgements

Ritambhara Singh
Jack Lanchantin
Arshdeep Sekhon

UVA Department of Biochemistry and Molecular Genetics: Dr. Mazhar Adli
Thank you