Deep Learning for Linking Epigenetic Modification and Gene Expression

Yixuan Li
yl3803@columbia.edu

Jason Wang
jason.wang@columbia.edu

Abstract

We adapted a recurrent neural network model to predict differential gene expression across different cell types. We planned to predict differential gene expression using single-cell DNA methylation and chromatin accessibility data. However, due to the current limitations of single-cell sequencing techniques, the dataset is too small and extremely sparse to train a valid model. To obtain meaningful results, we incorporated bulk sample histone acetylation data with histone methylation data as input to predict differential gene expression across cell types. Our model with more elaborated architecture and additional feature did not outperform the original model, although the difference is not significant.

1 Introduction

The human body contains hundreds of different cell types. Each cell has the same set of DNA sequences, but their functions can be drastically different due to various gene regulation mechanisms. Epigenetics is the study of all meiotically and mitotically heritable changes in gene expression that are not encoded in the DNA sequence itself. DNA methylation, RNA-associated silencing and histone modification are the most common ways to initiate epigenetic silencing and affect gene expression without modifying DNA sequences (Egger et al. [2004]). However, the exact mechanisms of gene expression regulation across multiple molecular layers are still not clear.

Recent progress in deep learning has been proven extremely successful in various prediction tasks, such as speech recognition, image recognition, object detection and other domains including genomics (LeCun et al. [2015]). Deep learning models are able to make extremely accurate predictions because their ability to capture intricate structures in large datasets, but the interpretability of such model is often poor. Algorithms such as DeepLIFT (Shrikumar et al.) and Intergrated Gradients (Sundararajan et al.) have made it possible to better understand deep learning models and extract rules. We propose that, using data from single-cell DNA methylation, chromatin accessbility and transcription, it is possible to develop a reliable model for gene expression prediction, which might in turn provide insight into the mechanism of gene expression regulation.

2 Background

2.1 DNA methylation

The methylation of C^5 position of cytosine residue has long been established as an epigenetic silencing mechanism (Holliday and Pugh [1975]). One of the most notable roles of DNA methylation is the transcriptional repression of certain genes. Methylated cytosines are known to be highly mutagenic, which will cause C:G base pairs to transition into T:A and suppress the CpG methyl-accepter site. Methylated DNA can either physically impede the binding of transcription factors, or bind to methyl-CpG binding proteins, which in turn recruits additional proteins such as histone deacetylase to the locus, causing further epigenetic modifications.

2.2 Histone modification and chromatin accessibility

Histone modification is another important aspect of epigenetic modifications. Eukaryotic chromatin is tightly packed into an array of nucleosomes, each consisting of a histone octamer core wrapped around by DNA and separated by linker DNA. The nucleosomal core, consisting four histone proteins, can be covalently modified or replaced by histone variants (Tsompana and Buck [2014]) These modifications, including histone methylation, histone acetylation, histone phosphorylation, can change the positions and availability of nucleosomes throughout genome, which controls the *in vivo* availability of binding sites to transcription factors and other regulatory proteins, affecting processes such as DNA replication and gene expression (Radman-Livaja and Rando [2010]).

2.3 Correlations between histone modification and DNA methylation

DNA methylation has long been associated with histone modification in their roles of regulating gene expression (Thurman et al. [2012]). Experimental evidence has shown a strong correlation between DNA methylation and histone modification activities. For example, H3-K9 methylation has been shown to be a prerequisite for DNA methylation (Lehnertz et al. [2003]), while DNA methylation can also trigger H3-K9 methylation (Johnson et al. [2002]). It is reasonable to assume that DNA methylation and histone modification work concurrently when regulating gene expression, although the exact mechanisms still remain to be elucidated. We envision that a reliable model to predict gene expression using both DNA methylation and histone modification as inputs might be able to provide insights into the mechanisms of gene expression regulation.

2.4 Single-cell epigenetic and transcription profiling

In order to understand the regulatory associations between the epigenome and the transcriptome, simultaneous profiling of multiple molecular layers is needed. Previously, the majority of such analyses are limited to bulk assays and ensembles of cells (Degner et al. [2012]) With recent advances in single-cell technologies, it is now possible to utilize variations between single cells to determine regulatory associations between molecular layers. Various protocols have been established, where the methylome and the trancriptome, or the methylome and chromatin accessibility can be assayed in the same cell (Angermueller et al. [2016], Pott [2017]). One protocol has been reported (Clark et al. [2018]), where parallel profiling of chromatin accessibility, DNA methylation and transcriptome has been achieved. We anticipate that this assay can be combined with deep learning techiniques to develop a reliable model for gene expression prediction using both DNA methylation and chromatin accessibility as inputs.

3 Methods

3.1 Data Collection

Single-cell data is obtained from scMT-seq Angermueller et al. [2016] and scNMT-seq Clark et al. [2018]. scMT-seq is a method for parallel single–cell genome–wide methylome and transcriptome sequencing and scNMT-seq is a single-cell method for parallel chromatin accessibility, DNA methylation and transcriptome profiling. Both of the data are available in the Gene Expression Omnibus under accession and GSE74535 and GSE109262.

Bulk sample histone methylation and acetylation data were downloaded from Roadmap Epigenomics Project (REMC). We collected seven different histone modification data (H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K9ac, H3K27ac) for five different cell types (R ID:)

3.2 Model Construction

We plan to construct the model based on a pevious work from DeepDiff, a model to interpret how dependencies among histone modifications control the differential patterns of gene regulation across different cell types Sekhon et al. [2018]. DeepDiff uses a hierarchy of multiple Long Short-Term Memory (LSTM) modules to embed the spatial structure of input signals and to model how various histone modifications cooperate. Attention weights are added before embedding layers to interpret the importance of different features. We plan to improve the model by considering two alternative

models, Pervasive Attention Elbayad et al. [2018] and Gated Recurrent Neural Networks Chung et al. [2014]. GRU is more computationally efficient as it controls the flow of information like the LSTM unit, but without having to use a memory unit. Pervasive Attention relies on a single 2D convolutional neural network with causal convolution that can outperform both RNN/LSTM and Attention based models.

3.3 Data Processing

Single-cell sequencing data are usually sparse compared to bulk sample data. Deep learning models often struggle with such sparse data, especially for models like LSTM or GRU which heavily rely on dense sequential attributes in order to learn a meaningful model. We attempted to solve this problem in a number of ways: 1) over-sample the positive examples (or downsample the negative examples); 2) pre-train our models with bulk data and optimize it with single-cell data. Each cell will be treated as a channel when fed into the model as inputs.

Our initial attempt was to use scMT-seq and scNMT-seq to predict gene expression; however, because of the data format, we were facing challenges in understanding and correctly processing the data into read counts with the formats which the model accepts. Because of the sparsity of the data, we also expected a unsatisfying performance of the model training with single-cell sequencing. Therefore, we modified our original plan to testing whether adding acetylation to the original histone modification that DeepDiff (Sekhon et al. [2018]) used could improve the model's performance in predicting differential gene expression.

4 Preliminary Results

4.1 Data Preprocessing

We downloaded the scMT-seq data from Gene Expression Omnibus (GEO). For each cell, it contains methylation rate of methylated region and expression rate of gene and therefore we could use methylation rate information to predict the expression rate for a single cell.

We first replace "NaN" values of methylation rate with 1.00. And observing that due to multiple experiements, for the same cell at the same methylated region, there are different methylation rates, we merge the different rows with the same cell name and methylated region and replace the methylation rates with the mean of them. We then construct our training data by transforming the current data frame. In the new input data frame, each row is a single cell, each column is the methylated region displayed as a tuple with the first value being the starting position and second value being the end position and each entry is the methylation rate of the cell at the current methylated region. The label matrix has the same shape as the input matrix. Each entry of the label matrix is the expression rate of the gene for a single cell at the methylated region.

4.2 Model Training

We started with improving the DeepDiff model training with bulk data. The data is from Roadmap Epigenomics Project (REMC) database and it contains ten different pairs of cell types. The genes are divided into 3 separate sets for training (10,000 genes), validation (2360 genes) and testing (6100 genes).

We used the model implemented in the original paper Sekhon et al. [2018] as our baseline. The original model used multiple levels of 32-unit LSTM networks to embed the histone modification data, in order to catch their long-term dependencies and predict the differences between gene expression across two different cell types. We optimized this model using the bulk-sample data as provided in the paper (Figure 1): with the same hyperparameters, GRU performs better than both LSTM and Elman-type RNN networks; 2-layer 256-unit GRU consistently offers better performance. The performance of this model is evaluated as Pearson Correlation Coefficient (PCC). Our best model performs gives PCC of 0.75, while the baseline model in the paper gives PCC of 0.60.

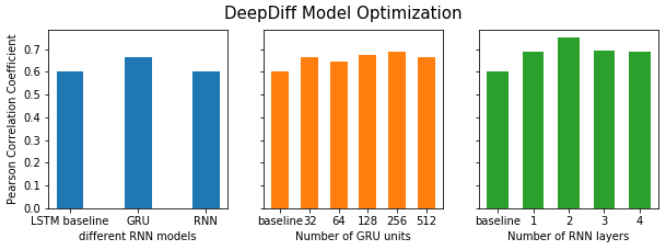
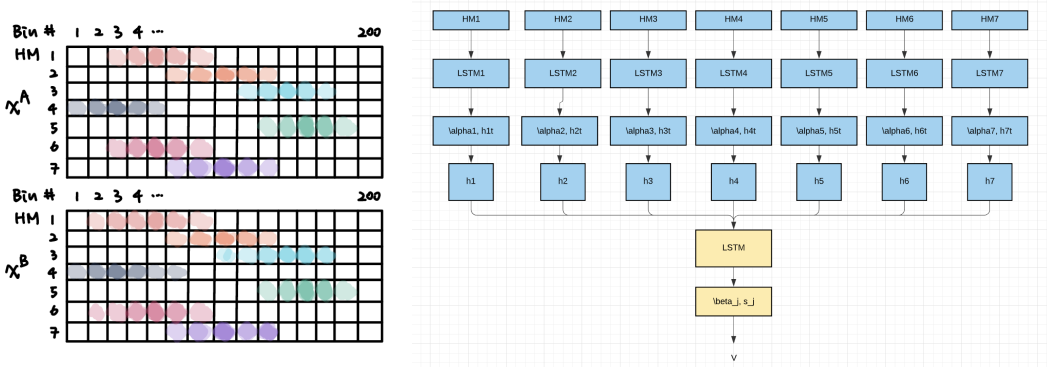


Figure 1: Optimization of DeepDiff model using bulk data.



(a) input format

(b) two-level embedding

5 Final Result

5.1 Data Processing

We downloaded histone methylation and acetylation data from ROADMAP database (REMC) for 18460 genes in total. And we focused on 5 cell types (E003, E005, E006, E116, E123) and 7 histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K9ac, H3K27ac). The newly added histone modifications are H3K9ac and H3K27ac. We used *"bedtools bedtobam"* to convert data from .tagAlign format to .bam files. And then we created a .bed file containing the genes that were used in the original paper, and selected bins of length 100 base-pairs (bp) from regions (+/- 10000 bp) flanking the transcription start site (TSS) of each gene. Finally, we used *"bedtools multicov"* to get the read counts and the signal value of all seven selected histone modifications from REMC in bins forms input matrix X, while log fold change in gene expression is the output y. We divided the genes into 3 separate sets for training(10,000 genes), validation(2360 genes) and testing(6100 genes).

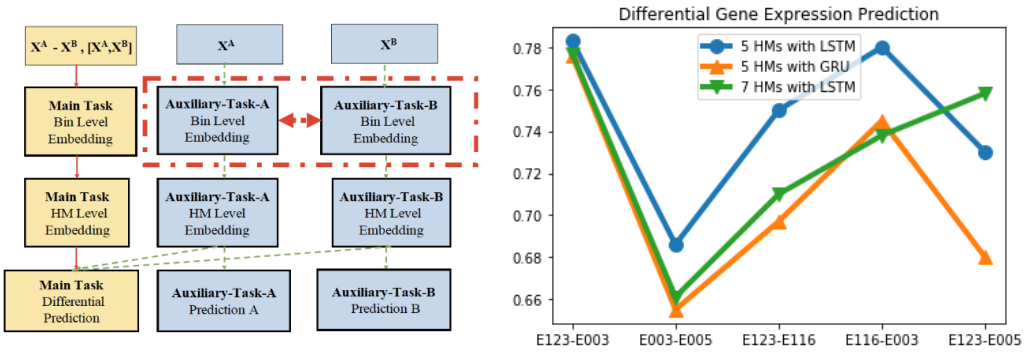
5.2 Model Architecture

5.2.1 Input

Given two cell-types A and B, and a gene g under consideration, the HM profile of gene g in A and B is denoted as X^A and X^B, respectively. As shown in left figure above, we consider $M = 7$ HM marks for each gene. Each HM signal across the $T = 200$ bins in cell-type A is represented by a row vector in X^A. Similarly, for cell-type B, each HM signal is represented by a row vector in X^B. In addition to the raw HM signals from the two cell types under consideration, we use difference and concatenation of the raw HM signals.

5.2.2 Two Level Embedding

There is a two level embedding of the model: *Level I Embedding* and *Level II Embedding*. The Level I Embedding module consists of a bin-level LSTM for learning the embedding of every HM,



(a) Auxillary Tasks

(b) Training Result

followed by a bin level attention mechanism. Bin level attention (α) reflects the relative importance of each bin or genome coordinate position for prediction. To efficiently represent the combinatorial dependencies between the various HMs, the model uses another LSTM as a second level embedding module. HM-level attention(β) reflects the relative importance of each HM for prediction.

5.2.3 Auxillary Tasks

The model also has two related tasks as auxiliary tasks for multi-task learning with our DeepDiff main task: *Cell-Specific Auxiliary – (Auxiliary-Task-A + Auxiliary-Task-B)* and *Siamese auxiliary loss term*. (fig. 3) In cell-specific auxiliary task, for auxiliary-task-A, X^A is passed through two levels of embedding and attention module specific for cell-type A. And similarly, for auxiliary-task-B, X^B is also passed through the two levels of embedding and attention module. The output embedding of the Level II Embedding unit v_A is passed through an multi-layer perceptron, which learns to map X^A to the target value for the gene in cell-type A. Similarly, v_B is passed through another multi-layer perceptron for cell-type B specific gene expression prediction. And experiments have shown that by jointly training cell-type specific gene expression with differential gene expression, the DeepDiff has improved performance. On the other hand, tshe siamese auxiliary loss term encourages the model to learn embeddings whose neighborhood structures in the model representation space are more consistent with the differential gene expression pattern.

5.3 Model Training

We trained our model with the auxiliary task and Siamese loss to improve our predictions. Observing that GRU produced a better result on the sample data, we replaced the LSTM model in the two-level embedding module with GRU to examine the change in performance. We trained our model with 30 epochs and a learning rate of 0.0001. Figure 4 shows the Pearson Correlation Coefficient (PCC) value for all DeepDiff variations for each cell-type pair(x-axis). Although GRU had a better performance on the sample data, it generates a lower PCC value when training with large cell dataset. And also, the model trained with 7 histone modification has a slightly lower score compared with the model trained with 5 histone modifications.

Training Result					
Method	E123-E003	E003-E005	E123-E116	E116-E003	E123-E005
LSTM with 5 HM	**0.783**	**0.686**	**0.75**	**0.78**	0.73
GRU with 5 HM	0.776	0.655	0.697	0.745	0.68
LSTM with 7 HM	0.777	0.661	0.710	0.738	**0.758**

6 Discussion

The model with GRU was expected to perform better than LSTM based on the result on the sample data. However, as shown in the table above, GRU has a lower performance across all five cell pairs. The sample data was provided along with the DeepDiff project and we don't know exactly how they sampled data and which cell types that they belong to. Therefore, it could be an outlier pair that

GRU uniquely performs better. Also, due to the hardware limitation, we only trained our model with 50 epochs. Since we don't know the exact training parameters of the model mentioned in the paper, we might failed to train our models with enough epochs or a proper learning rate or batch size. Similarly, it could be the same problem for models with 7 HMs. The fact that LSTM model with 7 HMs does not produce a significant difference to the original architecture proves the effectiveness of using histone modification for predicting differential gene expression. However, the lower result of our models raises the question of different effects of methylation and acetylation in gene regulation.

For future plans, We could try to incorporate additional epigenomic signals that may relate to differential gene expression and examine DeepDiff's performance. We would also like to explore different ways to interpret and validate the attention weights both on the current dataset or other epigenomic signals datasets. We believe that deep neural networks enhance our understanding of gene regulation by HMs and provide insights into principles of gene regulation through epigenetic factors.

References

C. Angermueller, S. J. Clark, H. J. Lee, I. C. Macaulay, M. J. Teng, T. X. Hu, F. Krueger, S. A. Smallwood, C. P. Ponting, T. Voet, G. Kelsey, O. Stegle, and W. Reik. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*, 13(3):229–232, 2016.

J. Chung, Çaglar Gülçehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555, 2014.

S. J. Clark, R. Argelaguet, C.-A. Kapourani, T. M. Stubbs, H. J. Lee, C. Alda-Catalinas, F. Krueger, G. Sanguinetti, G. Kelsey, J. C. Marioni, O. Stegle, and W. Reik. scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nat. Commun.*, 9:781, 2018.

J. F. Degner, A. A. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. D. Leon, K. Michelini, N. Lewellen, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard. DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482:390–394, 2012.

G. Egger, G. Liang, A. Aparicio, and P. A. Jones. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429:457–463, 2004.

M. Elbayad, L. Besacier, and J. Verbeek. Pervasive attention: 2d convolutional neural networks for sequence-to-sequence prediction. 2018.

R. Holliday and J. E. Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, 1975.

L. M. Johnson, X. Cao, and S. E. Jacobsen. Interplay between two epigenetic marks: Dna methylation and histone h3 lysine 9 methylation. *Curr. Bio.*, 12(16):1360–1367, 2002.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

B. Lehnertz, Y. Ueda, A. A. H. A. Derijck, U. Braunschweig, L. Perez-Burgos, S. Kubicek, TaipingChen, E. Li, T. Jenuwein, and A. H. F. M. Peters. Suv39h-mediated histone h3 lysine 9 methylation directs dna methylation to major satellite repeats at pericentric heterochromatin. *Curr. Bio.*, 13(14):1192–1200, 2003.

S. Pott. Simultaneous measurement of chromatin accessibility, dna methylation, and nucleosome phasing in single cells. *eLife*, 6:e23203, 2017.

M. Radman-Livaja and O. J. Rando. Nucleosome positioning: How is it established, and why does it matter? *Developmental Biology*, 339:258–266, 2010.

A. Sekhon, R. Singh, and Y. Qi. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics*, 34(17):i891–i900, 09 2018.

A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. *arXiv*, page 1704.02685.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. *arXiv*, page 1703.01365.

R. E. Thurman, E. Rynes, R. Humbert, and et al. The accessible chromatin landscape of the human genome. *Nature*, 489:75–82, 2012.

M. Tsompana and M. J. Buck. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin*, 7:33, 2014.