Predicting Gene Expression Using Epigenetic Markers on the Genome

CS230 Deep Learning

Sunil Bodapati – bodapati@stanford.edu, Timothy Daley – tdaley@stanford.edu, Sonja Johnson-Yu – sonjyu@stanford.edu

Overview

We use a 4-layer neural network to predict normalized gene expression from normalized chromatin openness.

This project builds off of previous work by Duren et al., which used a simple LASSO regression and yielded an R^2 ≈ 0.8 [1].

Openness dataset (X) shape:
200,000 genetic regions x 201 cell types

Target gene expression (Y) shape:
17,794 genes x 201 cell types

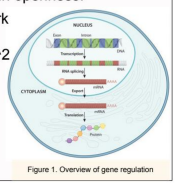

Figure 1. Overview of gene regulation

Dataset

- Obtained via ATAC-seq on 201 unique cell types
- Openness = proxy for possible epigenetic activity
- Openness signal in peaks

Preprocessing

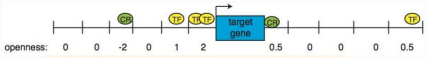

Figure 2. A proposed view of epigenetic regulation of gene expression. Transcription factors(TFs) and chromatin regulators (CRs) bind to DNA and affect gene expression.

Create Openness Bins

- Bin as function of distance from each gene's transcription start site (TSS)
- 2 x 1000 bins for 1 million BP (upstream and downstream)
- New shape: **17,794 genes x 201 cell types x 2000 openness bins**
- Distribution of gene expression approximately normal
- One outlier sample

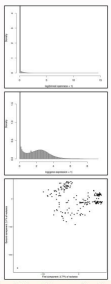

Figure 3. Top: histogram of log binned openness. Middle: histogram of log gene expression. Bottom: PCA projection of cells from gene expression data

Clustering Models and Results

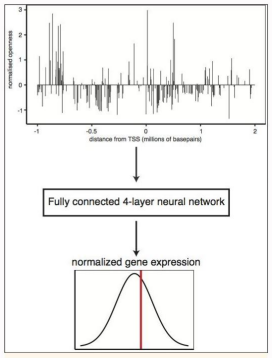

Figure 3. Overview of model pipeline

Experiments:
- Baseline LASSO
- Fully Connected model [1000, 1000, 1000, 1]
- L1 regularization
- L2 regularization
- Dropout
- Fully Connected model [1000, 500, 250, 1]

Training Parameters:
- lr = 0.001
- B1 = 0.9
- B2 = 0.999
- batch size = 10,000

Model Comparison

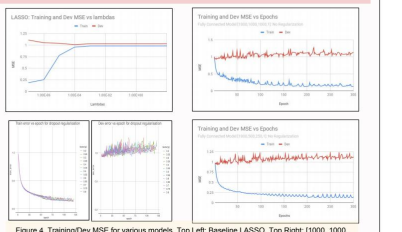

Figure 4. Training/Dev MSE for various models. Top Left: Baseline LASSO. Top Right: [1000, 1000, 1000, 1] FC. Bottom Left:Dropout (with varying p). Bottom Right: [1000, 500, 250, 1] FC

Error Analysis

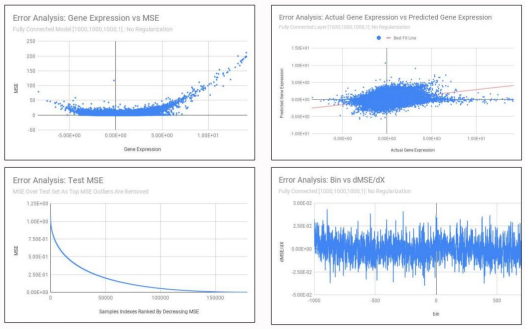

Figure 5. Top Left: Gene expression vs. MSE. Top Right: Actual gene expression vs. predicted gene expression. Bottom Left: MSE over test set as MSE outliers are removed. Bottom Right: Openness bin vs. dMSE/dX.

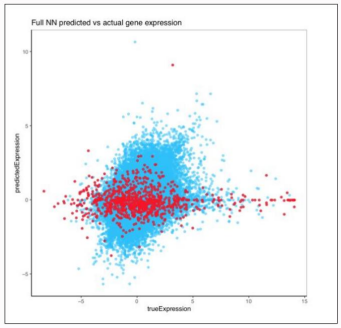

Figure 6. Highlighted outlier sample 91, actual gene expression vs. predicted gene expression

Error Summary

- Predictions skewed towards zero
- Greater prediction error for genes with higher expression
- Mean square error decreases exponentially as outliers removed
- Uniformly noisy distribution of bin contribution to MSE
- One cell type had high variance in prediction vs. true expression

Conclusion

- Lasso baseline dev MSE ≈ 1
- 4-layer FC overfit
  - train MSE ≈ 0.15, dev MSE ≈ 1
  - regularization ineffective
- A few samples cause the majority of the MSE
  - namely one cell type

Future Work

- Bucket phenotypically similar cells
  - reduce cell-specific effects that impact generalization
- Use sequence information

References

[1] Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proc Natl Acad Sci USA. 2017;114(25):E4914-E4923.