Generating Regulatory Sequence to Produce Target Expression

Nic Fishman,[1] Georgi Marinov,[1,2] Anshul Kundaje[1,2]

[1]Department of Computer Science, Stanford University
[2]Department of Genetics, Stanford University

Stanford Computer Science

Abstract

The abundance of high quality gene expression data afforded by the recent development of Massively Parallel Reporter Assays (MPRA) has created an abundance of data for developing a deeper understanding of transcription factor (TF) binding. Here we show that convolutional neural networks are capable of learning the motifs that underlie TF binding and predicting expression using these motifs at various amino acid concentrations [AA]. Using this result we develop a generative adversarial network that can build segments of regulatory sequence to produce specified gene expression at varying [AA].

We find that a combination of the MSE between the predicted expression and the target expression and the standard WGAN-GP loss give the best results for learning to produce sequence given target expression levels.

There is still more work to be done, especially in tuning the combination of the losses and in trying to resolve the diminishing return on longer training that is found for all values of lambda when generating sequence.

Stanford University

Genetic Background

Regulatory Sequence

- Instructions for how much protein to make
- Encoded in discrete strings of basepairs called motifs
- Complexity comes from motif interactions, which depend on number and position of motifs

MPRA

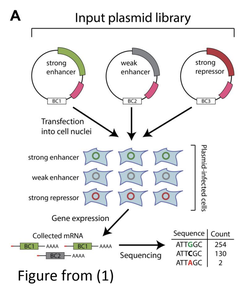
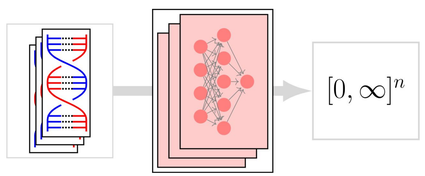


Figure from (1)

- MPRA allows testing expression of thousands of regulatory sequences at once
- Combinatorially combining motifs in regulatory sequence and get associated expression allows decoding of the lexical grammar governing expression (2)

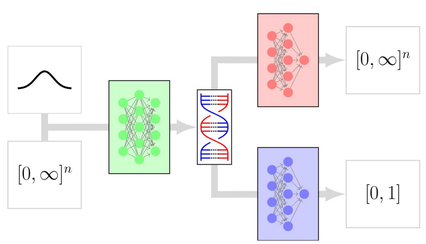
Predicting

Predicting Expression



Predicting protein expression from regulatory sequences.

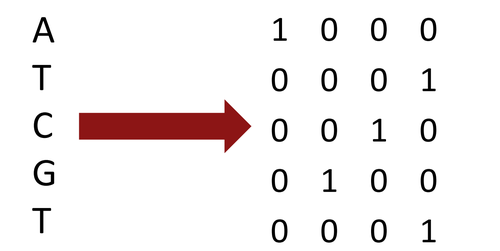
Targeted Sequence Generation



Given a target expression level generate sequence predicted to produce that expression level.

Data and Data Processing

The data comes from MPRA experiments (2), which produce the DNA sequence (ACGT) and corresponding protein expression. Each row is a sequence and the corresponding mean expression under several conditions.



The only feature work is to one-hot encode the alphabetic sequences.(1)

Models

Predicting Expression

Random Architecture Search

Architecture Property	Distribution Drawn From
Number Convolutional Layers	$\sim Uni(2, 4)$
Filters per Convolutional Layer	$\sim Uni(5, 50)$
Filter Size	$\sim Uni(4, 15)$
Number Dense Layers	$\sim Uni(1, 5)$
Units per Dense Layer	$\sim Uni(5, 100)$
Regularize all Layers	$\sim Bern(p = 0.5)$

- Train several regressors
- Select best based on integrated gradients ratio of motif importance over total importance

Targeted Sequence Generation

WGAN GP Loss (3)

$$L = \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}} \sim \mathbb{P}_g}[D(\tilde{\boldsymbol{x}})] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_r}[D(\boldsymbol{x})]}_{\text{Original critic loss}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}}\left[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2\right]}_{\text{Our gradient penalty}}.$$

Regressor Loss

$$L_R = \frac{1}{m}\sum_{i=1}^{m}(Y - \hat{Y})^2$$

Overall Loss

$$\mathcal{L} = \lambda L_D + (1 - \lambda) L_R$$

Evaluating Generated Sequences

- LOO accuracy 1-NN in learned feature space (4)
- Predicting expression via ensemble of regressors
- Motif identification and frequency analysis

Results Overview

Predicting Expression

Trained for 1000 epochs, with early stopping.

Model Rank	Training MSE	Test MSE	Motif Importance
1	0.014043	0.026653	1087.418475
2	0.002005	0.028467	1084.981529
3	0.008275	0.030267	1177.912115

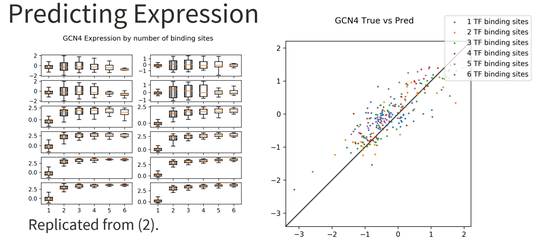
Targeted Sequence Generation

Trained for 5000 epochs.

	Metric	Optimal/Real Value	Generated
$\lambda = 1.0$	1-NN LOO	0.5	0.65
	Motif Count Per Sequence	17.54	14.114
	Predicted Expression MSE	0.0	0.801
$\lambda = 0.5$	Metric	Optimal/Real Value	Generated
	1-NN LOO	0.5	0.77
	Motif Count Per Sequence	17.54	11.894
	Predicted Expression MSE	0.0	0.205
$\lambda = 0.0$	Metric	Optimal/Real Value	Generated
	1-NN LOO	0.5	1.0
	Motif Count Per Sequence	17.54	8.336
	Predicted Expression MSE	0.0	0.792

Prediction Results

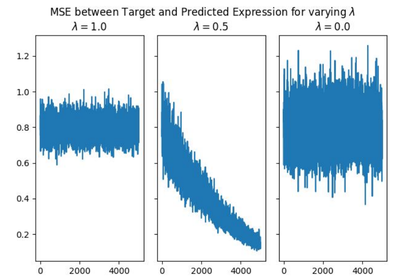
Predicting Expression



Replicated from (2).

Motif Importance Scoring

Generation Results



Discussion

- The regression model is definitely successfully learning to identify motifs, which is a great sign.
- The generated sequence from the WGAN is relatively similar to the real distribution and backpropagating on the predicted expression leads to much better accuracy on achieving target loss. This fully validates the pipeline.
- In training the GAN there is an issue where training stops leading to improvements in the 1-NN and motif count metrics after the first few thousand epochs. It would be nice to try to understand why this happens.
- There is a tradeoff in GAN training between hitting the given target expression and producing "realistic" sequence. It would be good to try annealing the lambda term to see if this tradeoff can be resolved.

Citations

1. M. Rajiv, et al., "Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays" bioRxiv preprint bioRxiv:393926, 2018.
2. D. Van Dijk, et al., "Large-scale mapping of gene regulatory logic reveals context-dependent repression by transcriptional activators" *Genome Research*
3. Gulrajani, Ishaan, et al. "Improved training of wasserstein gans." Advances in Neural Information Processing Systems. 2017.
4. Borji, Ali. "Pros and Cons of GAN Evaluation Measures." arXiv preprint arXiv:1802.03446, 2018.