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 sequences.

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.

Generating Regulatory Sequence to Produce Target Expression

• 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

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

Results Overview

• The regression model is definitely the best to try annealing the lambda term to see if this tradeoff can be resolved.

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