CS 224d Project Proposal

Govinda Kamath and Jesse Zhang

28 April 2016

1 Problem Description

The genome of an organism is a long string of the nucleotides A, C, G, and T. The genome contains genes which are translated into proteins essential for life, and genes are regulated by a variety of epigenetic factors (e.g. transcription factors). Regions of the genome that are very far away from one another could influence the regulation of a particular gene. The purpose of this project is to explore how recurrent neural networks (in particular bidirectional LSTMs) can model and predict the long-range interactions within a genome.

2 Data

ChIP-seq and ATAC-seq are two assays used to obtain the sequences of a cell type that are active. Several datasets are available online (ENCODE), and these datasets have both sequences obtained from multiple experiments corresponding to the areas of the genome that are active. If a dataset consists of N sequences, we can obtain N training examples from the dataset. The input will be one of the N sequences, and the model will attempt to predict the other sequences. Each sequence will be broken into length-k substrings or k-mers (a sequence with k-mers is analogous to a sentence with words).

3 Methodology

The training and testing of the model will be as follows:

1. Map k-mers to vectors using methods such as word2vec or GloVe.
2. Train an LSTM using the training examples.
3. Cross validate to select model hyper-parameters such as network structure, activation function, and regularization strength.
4. Given a new sequence, determine which parts of the genome are most relevant.

4 Related work

While several works ([1], [2], [3], [4]) have used deep neural networks to predict how a sequence of nucleotides is regulated, none of them use a pure RNN architecture. Also, none of them attempt to predict long-range interactions within the genome.

5 Evaluation plan

We will evaluate our results by holding out some of the training data. We will then attempt to predict the relevant sequences given an input sequence. We expect that our model will fail to predict all of the

sequences and simultaneously predict incorrect sequences, and we will quantify this performance using false negative and false positive rates. We will compare the performance of this model to a simple softmax regression (predicting multiple labels given a new sequence).

References

[1] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 2015.

[2] David R Kelley, Jasper Snoek, and John Rinn. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *bioRxiv*, page 028399, 2015.

[3] Daniel Quang and Xiaohui Xie. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *bioRxiv*, page 032821, 2015.

[4] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, 12(10):931–934, 2015.