**Deep learning in computational biology project.**

By:

Daniel Yanovsky 314826975

Saar Aloni 316440023

**Goal:** Predict RNA binding intensity for each RNAcompete probe, for a protein that is given by DNA sequences.

**Input:** some files of DNA sequences that were observed in different concentrations, for a single protein, and a file of RNA sequences.

**Output:** binding intensities, for the given protein, of each of the RNA sequences that were given in the input.

**How does it work:**

First we draw the RBP sequences data, for each DNA sequence in the files we draw every 8 letters sequence from it using a ‘sliding window’.

For each of the 8 letters sequence we convert it to a matrix representing the DNA letter with a one hot vector. We get a 8 by 4 matrix for each sequence.

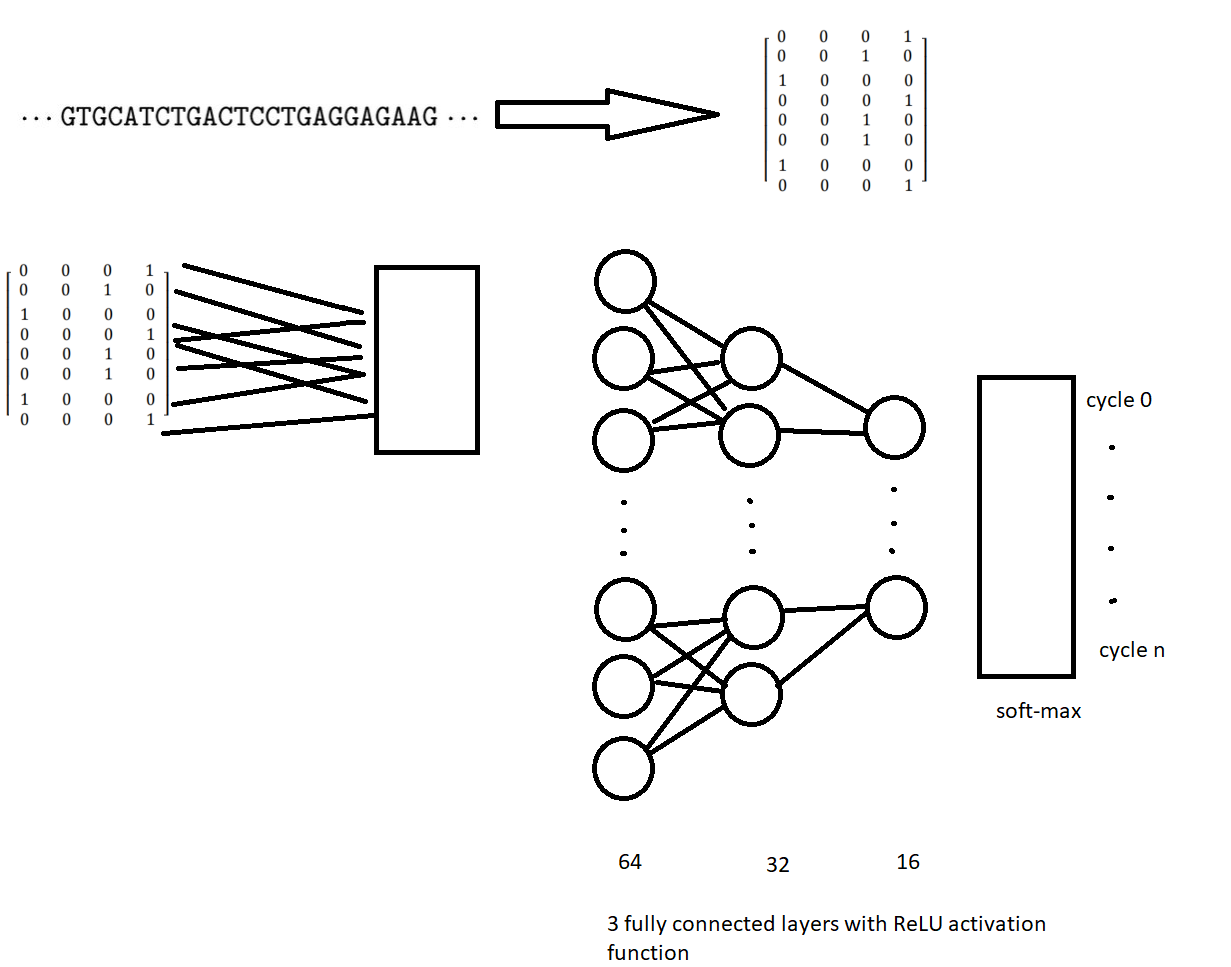
The now converted to matrix data is fed to Convolutional neural network, first we use kernels on the matrixes, kernels are sized 4 by 4, and we use 8 of them. The results are then flattens and fed to a fully connected network of neurons. There are 3 layers of neurons, the number of neurons are 64, 32 and last 16, the layers use ReLU activation function. After which there is a multi-classification softmax layer.

The optimizer we use is ‘adam’ with a mean squared loss.

The softmax layer uses every file from the DNA sequences that we receive in the input as a possible output of the layer.

After the data was fed into the network, we move to the RNA sequences. We preform the same operation to the sequences from the RNAcomplete file and draw every 8 letters sequence from all RNA sequences in the file.

We use the trained network to predict for each sequences which concentrations it fits, then we score intensity for each RNAcomplete sequence.



Results on training:

on training files by using pearson correlation with the given RBP files. We reached in average to about 0.20 correlation.

Resources:

Usage of memory stand at about 5Gb

Time takes about 5 minutes to train the network and another 5 minutes to predict the RNA scores.

CPU uses 30% of 3.6GHz CPU.

Result of traing file

|  |  |
| --- | --- |
| file | Pearson correlation with training intensity |
| RBP1 | 0.3 |
| RBP2 | 0.29 |
| RBP3 | 0.1 |
| RBP4 | 0.08 |
| RBP5 | 0.18 |
| RBP6 | 0.37 |
| RBP7 | -0.05 |
| RBP8 | 0.06 |
| RBP9 | 0.12 |
| RBP10 | 0.1 |
| RBP11 | 0.16 |
| RBP12 | 0.32 |
| RBP13 | 0.44 |
| RBP14 | 0.16 |
| RBP15 | 0.2 |
| RBP16 | 0.31 |