# Assignment 3

# Transcription Factor Binding Prediction Report

Xiaocong Li 92923958

Provide the labeled data of SP1 transcription factor binding and non-binding sites on human chromosome1. There are 1000 sequences for binding sites and 1000 sequences for non-binding sites. Each sequence has 14 nucleotide base pairs. There are four different nucleobase types in the DNA sequence: adenine (A), cytosine (C), guanine (G), thymine (T). The sequences in the dataset are also denoted by these letters.

The file “train.csv” contains three columns:

- id: The id of transcription factor binding and non-binding sites.

- sequence: The raw sequence data of transcription factor binding and non-binding sites.

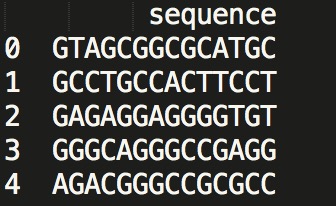
- label: Denotes whether the sequence is a transcription factor binding site or not. 1 is for TFBS and 0 is for non-TFBS.

“Test.csv” contains two columns:

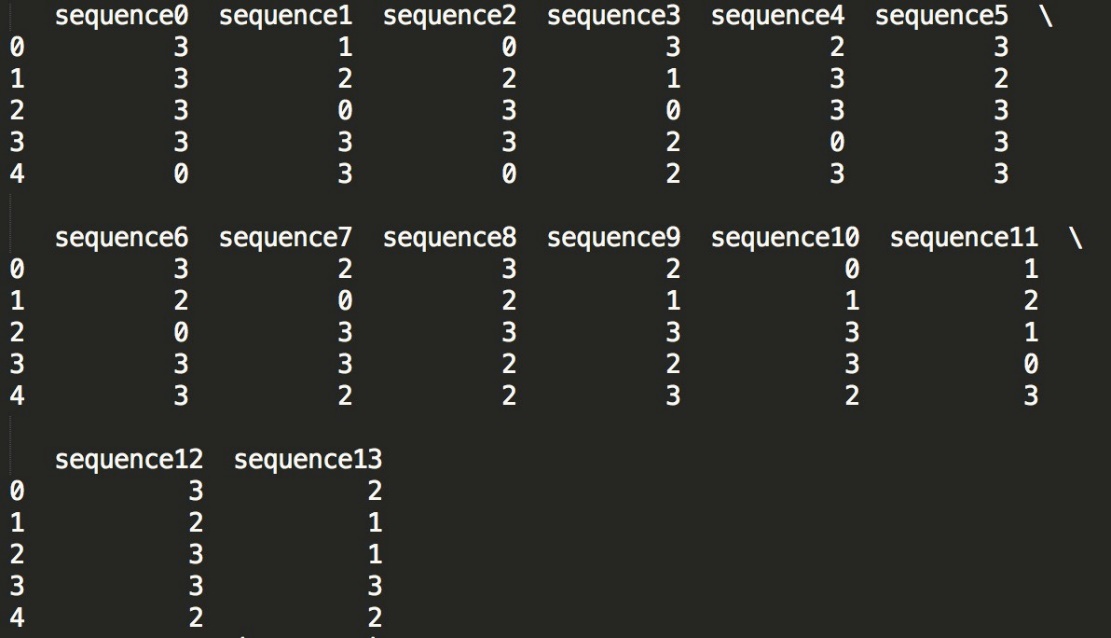
- id: The id of transcription factor binding and non-binding sites.

- sequence: The raw sequence data of transcription factor binding and non-binding sites.

# **1. Data Processing**

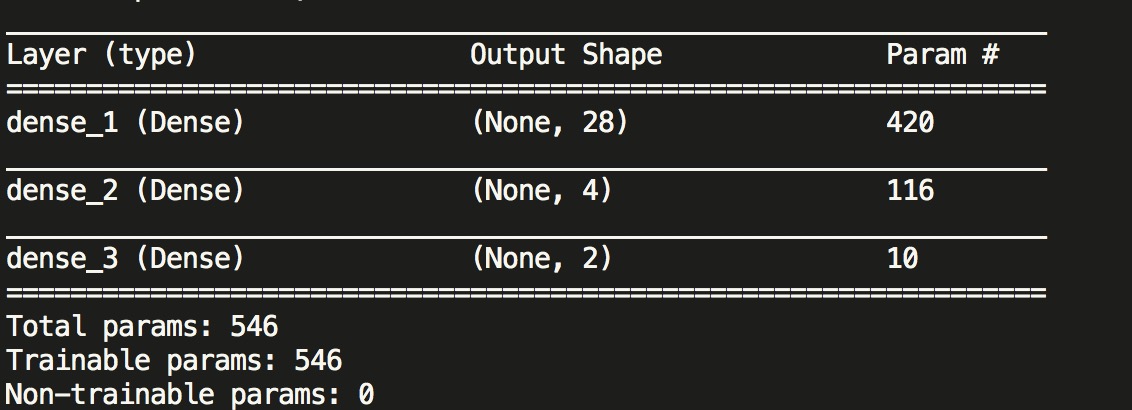


As shown above, we can find the length of a column of sequences is 14, consisting of A, T, C and G. Denote A, T, C and G as 0, 1, 2 and 3. Consider them as 14 rows and we can have a new picture. Shown as:



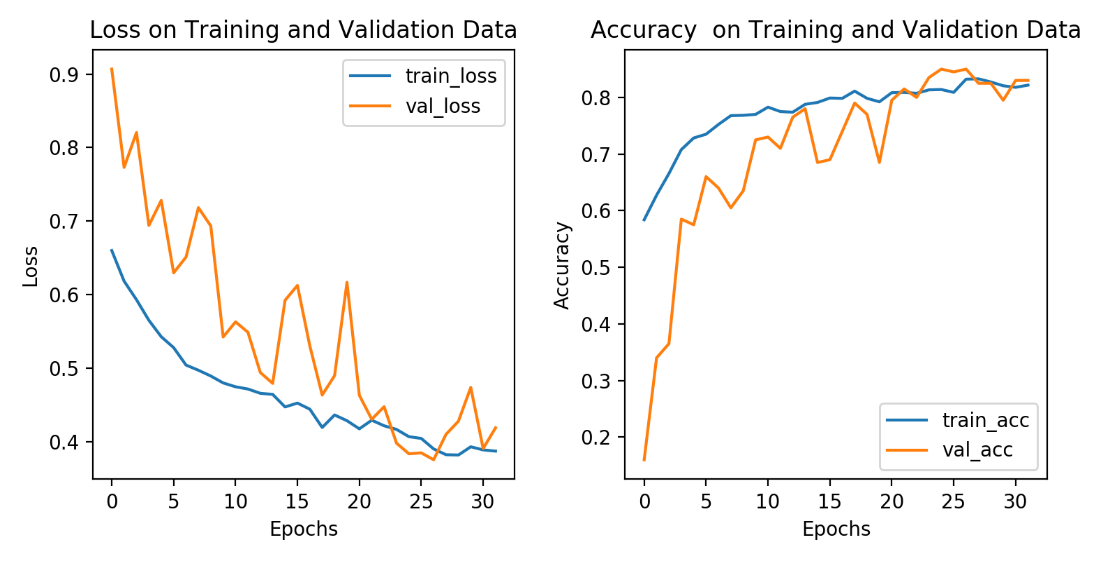
# **2. Neural Network Model**

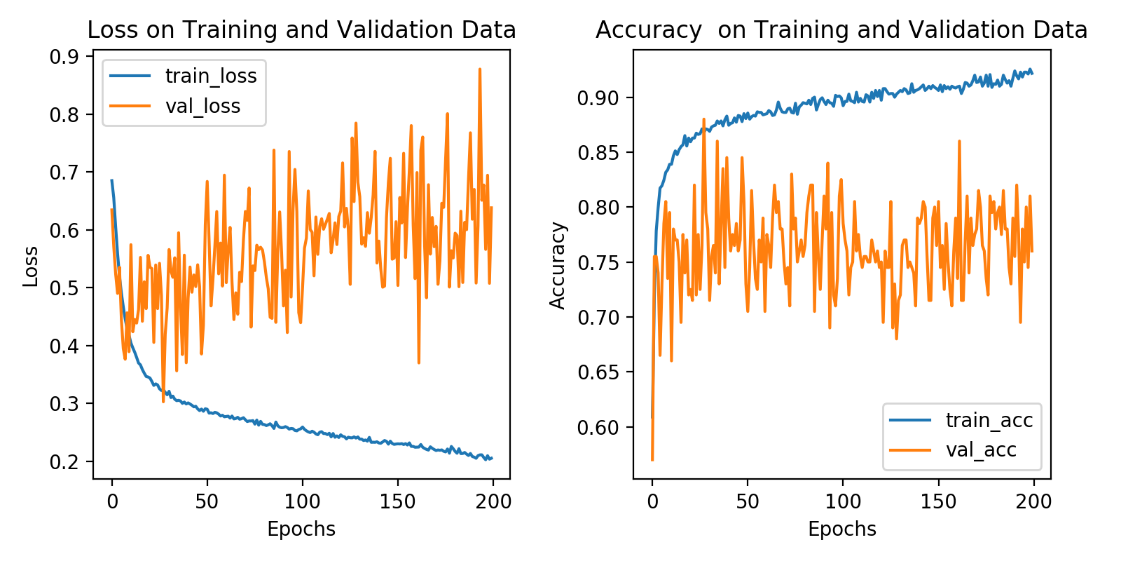
After data processing, now the data has a format as 1000 × 14, and each one ranges from 0 to 3, which can be imported into a neural network. Based on Keras, build a three-layer neural network. The parameters of this MLP than can be trained. The following picture shows the result:



# **3. Results**

During the first 30 iterations, we can conclude that the loss of training data decreases steadily while the loss of testing data varies a lot. Nevertheless, the decreasing trend is obvious. On the other hand, the accuracy of training data and that of test fit relatively well. Both increase.





However, after 200 times of iteration, the images of loss and accuracy show significant [indication](file:///C:\Users\muzij\AppData\Local\youdao\dict\Application\7.5.2.0\resultui\dict\?keyword=indication) of overfitting. At this time, strategies such as dropout are taken to control the iteration.

The score of the competition as team Lee is 0.81274 in private part and 0.77844 in public part.