Neural Network

1March2019

Agenda

I. Summary of Data Formatting
II. Bidirectional Experiments
III. Controls (Random sequences)
    A. Random Sequences
    B. Random PWM

Agenda

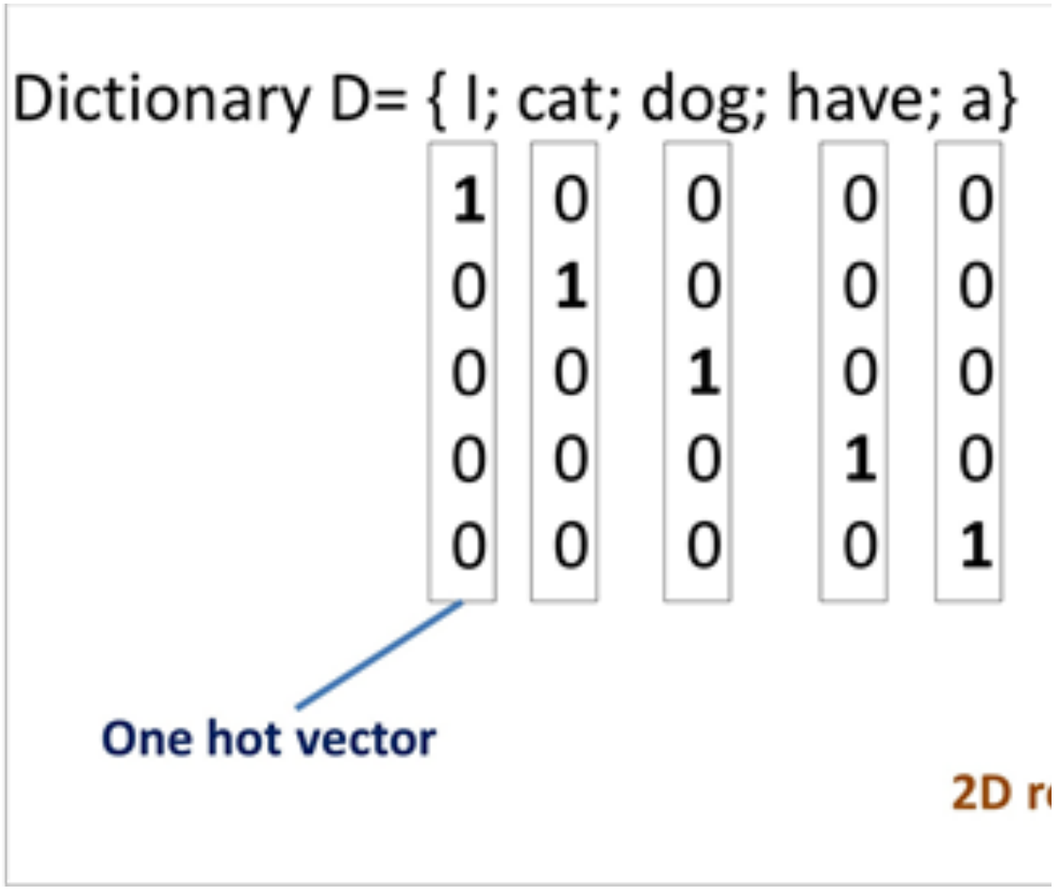
I. **Summary of Data Formatting**
II. Bidirectional Experiments
III. Controls (Random sequences)
    A. Random Sequences
    B. Random PWM

>VT44110|0|MEMB004B|-|2431
TAGGAACTCTTTTTCCCTTTTTCTCTCAGTGCAGCGATTGCTCTTGATGTCATAGTCGTTGGGCTGGT
CATTGTCATCACACGCCGGTTCATTCGTTGCATCATCACGCAACGAGGTGCGATCGATCCGCTATCA
GGCGAATGCAATTGCATCGGTGCAGTTTACGGCCTGATAAGCCCGGACTGCGGGGCTATGCGACAA
CTGCCCCTATTCCTGTTTTGCCCCTGACATTCGGTCGCCGGCGCATGCATTTATGTACGtgccatgcgcc
attagccatatgccatatagcctatatAGATGGTCGGACGTTGGGCTGTCCATAAACCATTTCAAATGCTATTTT
TGGATGTAGGAAATGGCCAAAGATGAGTTGCCAATTTTATTTGCATATGTATACAGGAAAATGTACCAA
ATTTGTTTATTAATAACGTCACAGATTGACTGAACAACTTTTAACGCCTAAGTTAAAGTTAAAAAAAGAA
GTTTATAAAAAGAAATTTCGAATATTTAAAGAATGTGCATTTCAAAATAATCCCACTAGGTCAACATGTA
TCCGGAAATACGAAAGCTAACAAAGACTATGATTTGTTTTTCTAAAAATAAAAATTATAATTTTAAAATCA
TCTACAAAGAGTTGACGATGAGATTAAAATGTTTTTTGAATTCGCAAAAAAAAATGTtttataatttaaaaattaa
agatttaaagtcacgaaaaactatgaatttaaaatataaaggattaaaaaattCCCAAGTTAAGTTATTATTCAAGTTTCTA
GCCAATTAACGTTTATTCATTTTTCATTAGCCAAAATTTAAGTTTTTATGTGTCCAATGGACATTTAATTG
GTTTTGTTCGGAACGACATTGACGAAGAACCAATTATGACTGAGCCTATGTAATATCTGAACACTCAA
ACAAACAATAGCTGTTATTAAAATCGCCATTTAATGTATTATTTAAGCCTTTTGACAAAGGGCACACAC
CCACCGATTATCTTATCTGTGTTGATATCATAAATTTTTGTTGTCAGGACTCTTACACAGTTGCAACTAT
TATTATTACGTTTTTTTCGATTTACACTTTCCACGAGAGTTTGGATTTAGTTAACTTAAACTAGTAAACAT
TACTAGTTAGTAAAATTTTAAAACGAATACCTTTAAAAACTTTAATCAACATAAAATTAATGTAAAATCCA
TTAGAAATTTAATAAAAACTATATTTAACACTCCCACTAACTTTTAGCCATCTAATCttattttatttatttatttttat
ttTCACACTTTGTTCGGCCTTTAATCCACTTGCAGCCGCTAGAGGGCGCGTTCGGCAATTACGGAGAA
AAGAGATCTCTTCAAGTGCATTTCTC

Across 24 Species

Link to file: https://github.com/DiscoveryDNA/TFBS_presence/blob/master/data/raw/outlier_rm_with_length_VT59000.fa

One Hot encoding nucleotide sequence: Using the entire nucleotide sequence as the backbone for TFBS scores. The nucleotide sequence is one hot encoded. Which means that each nucleotide (ATC or G) is a dictionary and each position is coded along that dictionary.



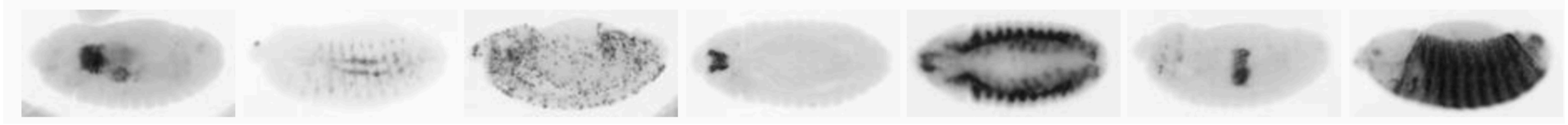

```
base_pairs = {'A': [1, 0, 0, 0],
'C': [0, 1, 0, 0],
'G': [0, 0, 1, 0],
'T': [0, 0, 0, 1],
'a': [1, 0, 0, 0],
'c': [0, 1, 0, 0],
'g': [0, 0, 1, 0],
't': [0, 0, 0, 1],
'n': [0, 0, 0, 0],
'N': [0, 0, 0, 0]}
```

-The entire sequence is flattend. For example, AGCT would be transformed into [1,0,0,0,0,0,1,0,0,1,0,0,0,0,0,1] where the first four represent A and the next four represent G and so on.

Classification: Each sequence is binary. The positive or negative presence of expression.

expression in early embryo:
1 = positive 0 = negative

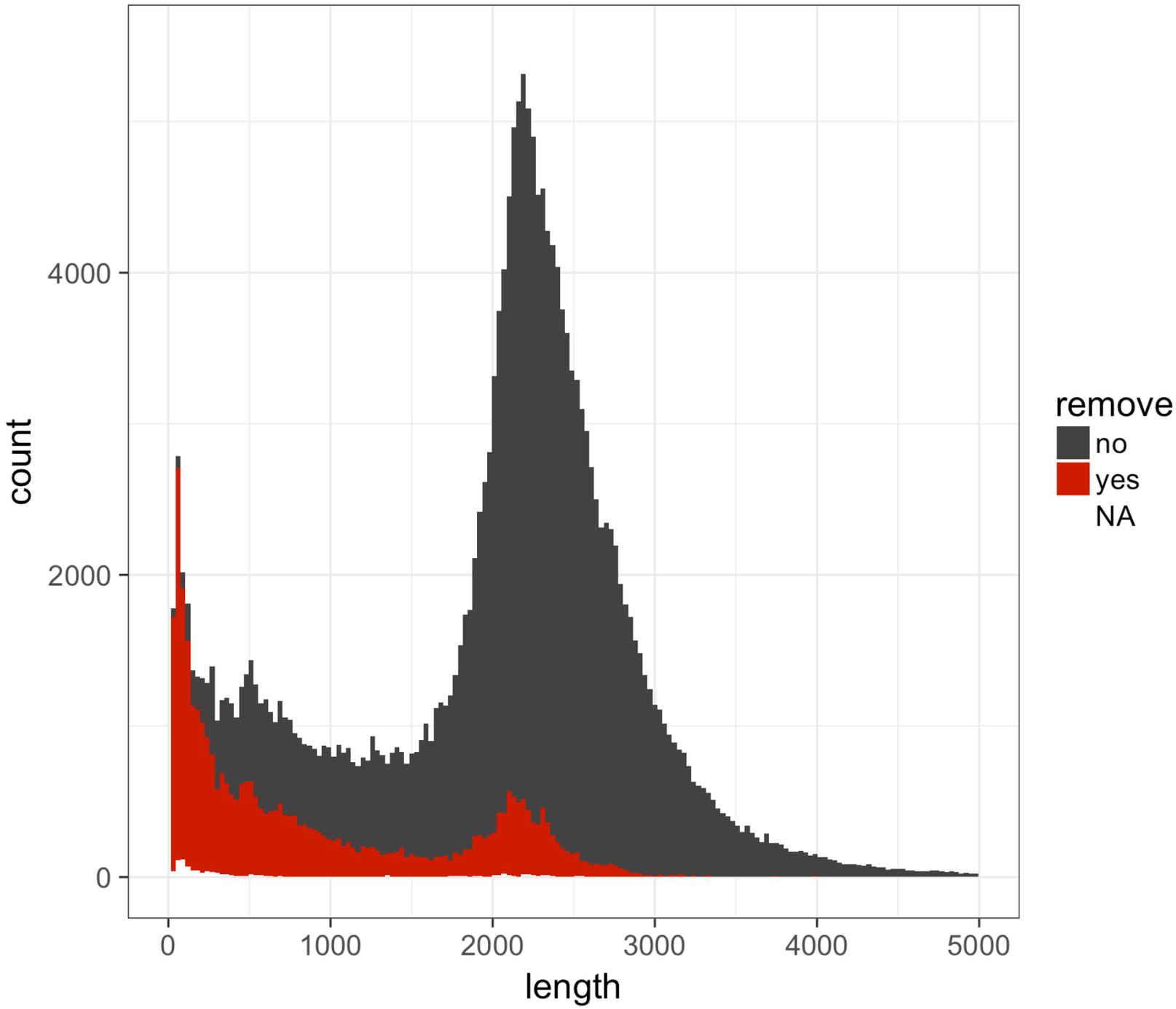
Link to actual data.



Sequence Length: They all need to be the same length, which they are not

The major things to note are the following:

- DNA sequences are of difference lengths, some very short (100~ bases), some very long (3000~ bases).
- Since most sequences are in the length range 1000 - 2000, we decide to only take the first 1000 bases of each sequence to train the neural network and make the predictions. If too long, simply truncate it to length 1000. If too short, simply fill with zeros to extend it.



Transcription Factor Binding Sites

So Far: Bicoid, Caudal, Eve, Zelda.

- [ ] Adam, do you know for sure which dataset you formatted the data from? Link: https://drive.google.com/drive/folders/19LV8QSPFbsEvglt785RUDKcxHoOiQ5rX

Essentially each position is tested if it is the start of a TFBS and a score is give to each letter position evaluating how sure a TFBS starts there.

EXAMPLE FAKE TFBS: AATTATAC

GTATAATTACTACAAATTATACTTTATTTATACAC

This is done on the positive and negative strands.

Each strand direction will have different scores

EXAMPLE FAKE TFBS: AATTATAC

GTATAATTACTACAAATTATACTTTATTTATACAC

GTGTATAAATAAAGTATAATTTGTAGTAATTATAC

Strand Specificity

We are subsetting for only the positive strand.
- To Do: Include the other direction.

Agenda

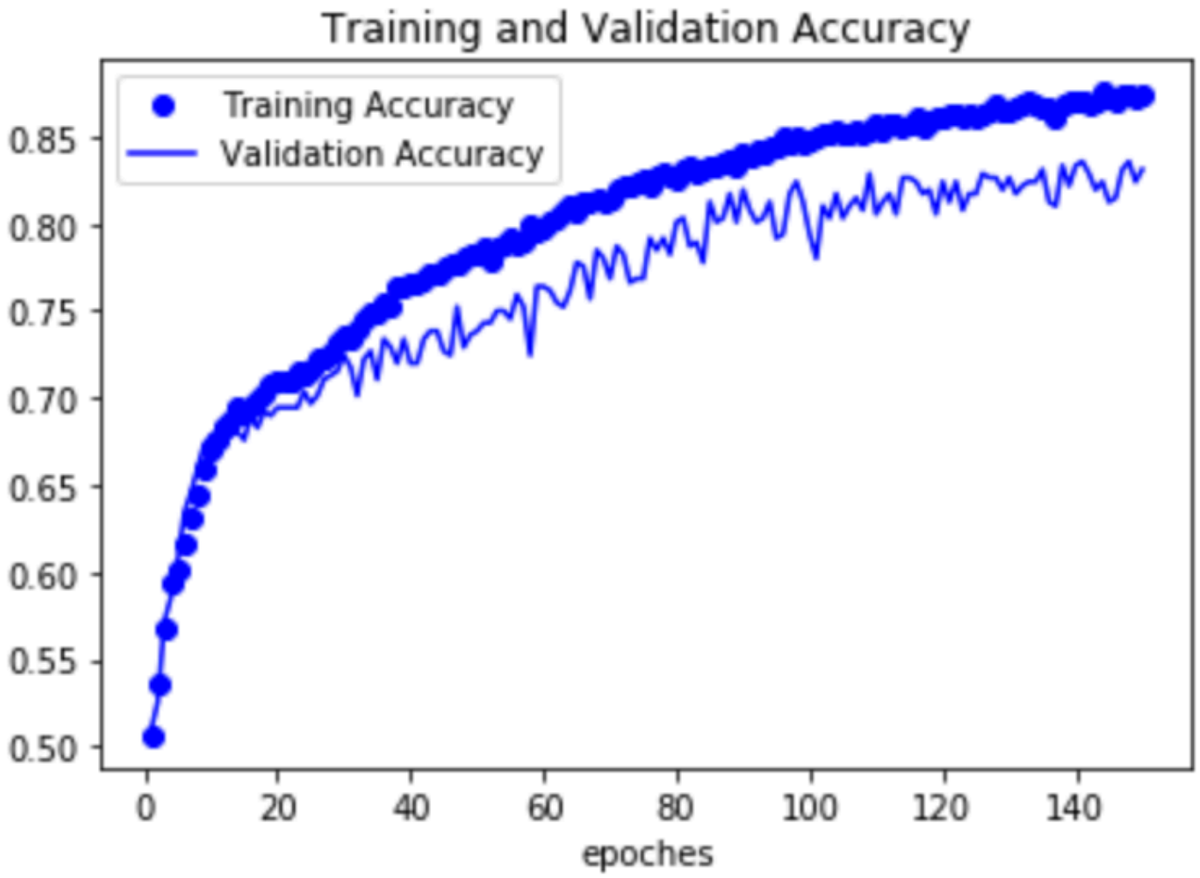
I. Summary of Data Formatting
II. Bidirectional Experiments
III. Controls (Random sequences)
    A. Random Sequences
    B. Random PWM

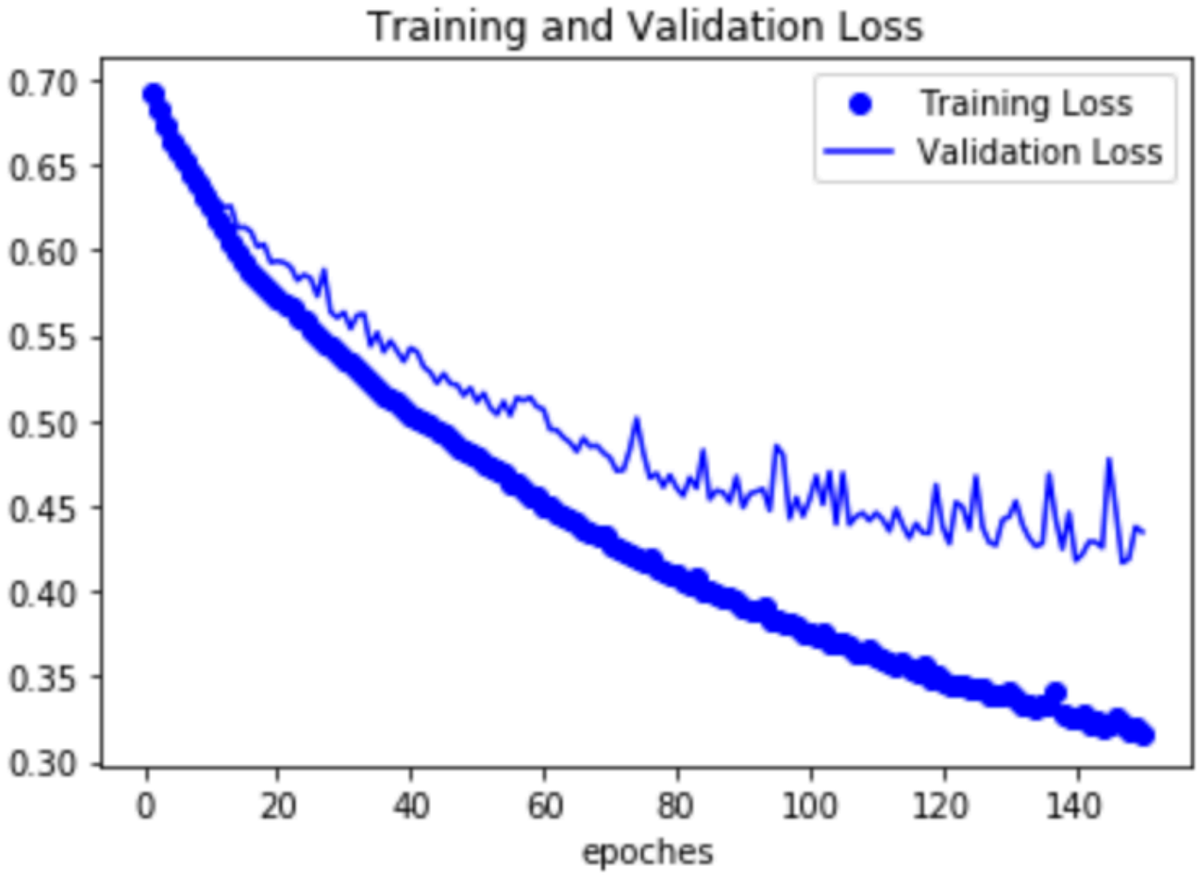
Bidirectional Architecture

[2019-01-29_shuffling_on_padding_at_the_end_with_size_1000_bidirectional.ipynb]

Accuracy jumped to over 85%

- using only a subset of the data
- Only 4 TFBS
- Only on the positive strand

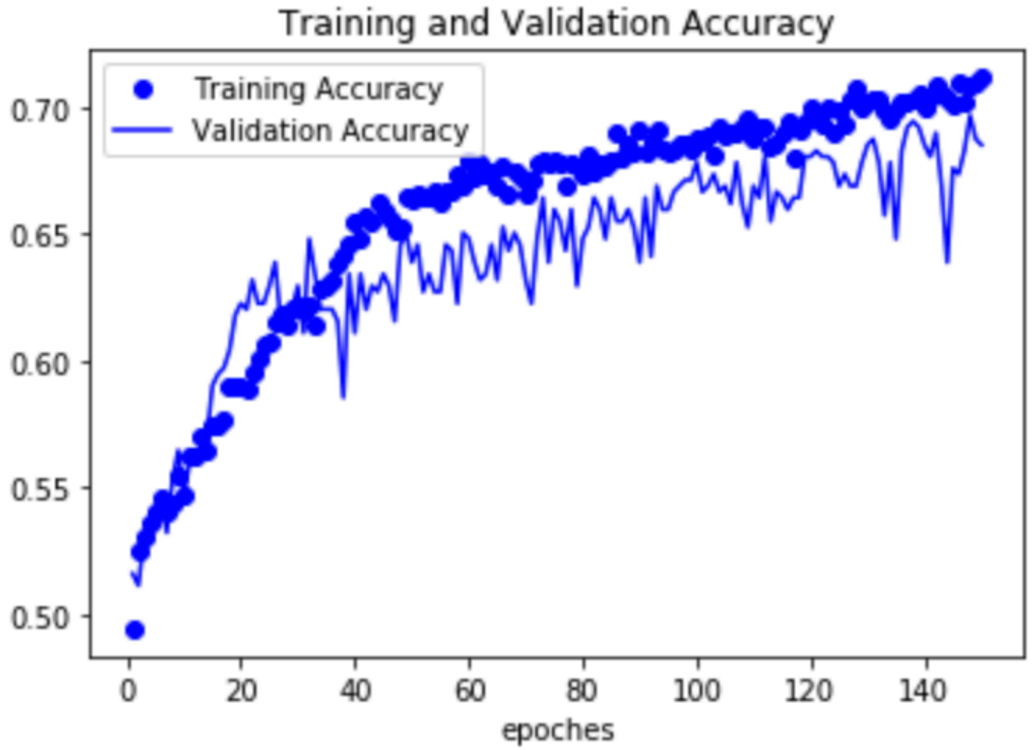


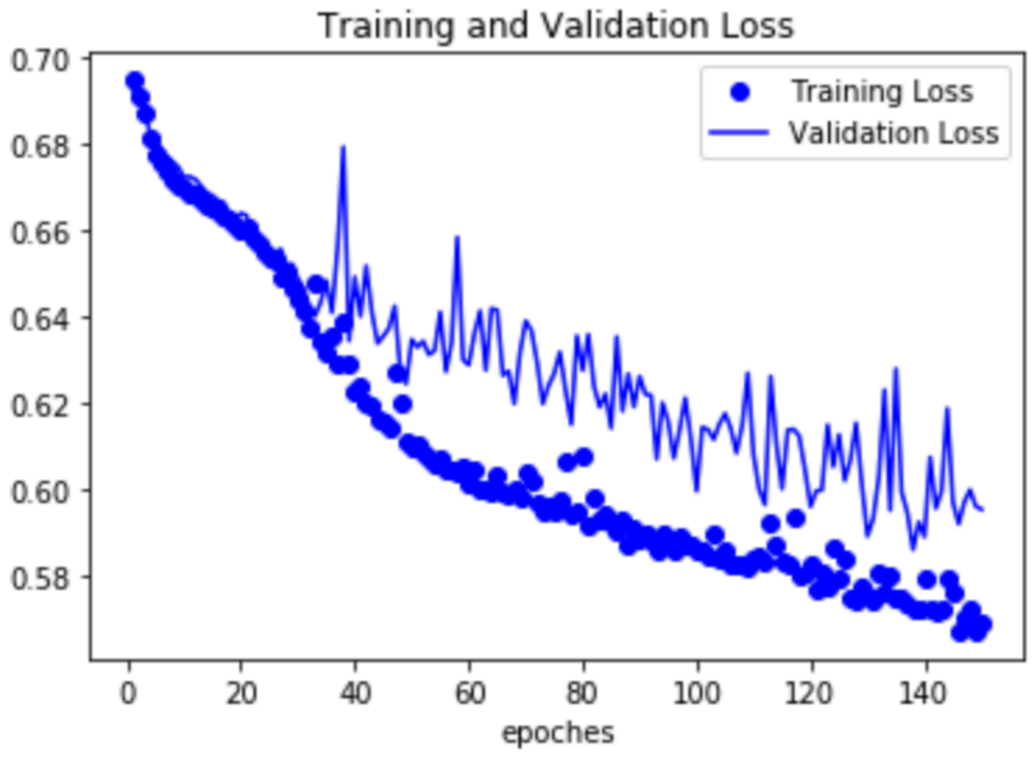


No TFBS, only nucleotide information

[2019-02-04_shuffling_on_padding_at_the_end_with_size_1000_bidirectional_no_TFBS.ipynb]

Accuracy 71%

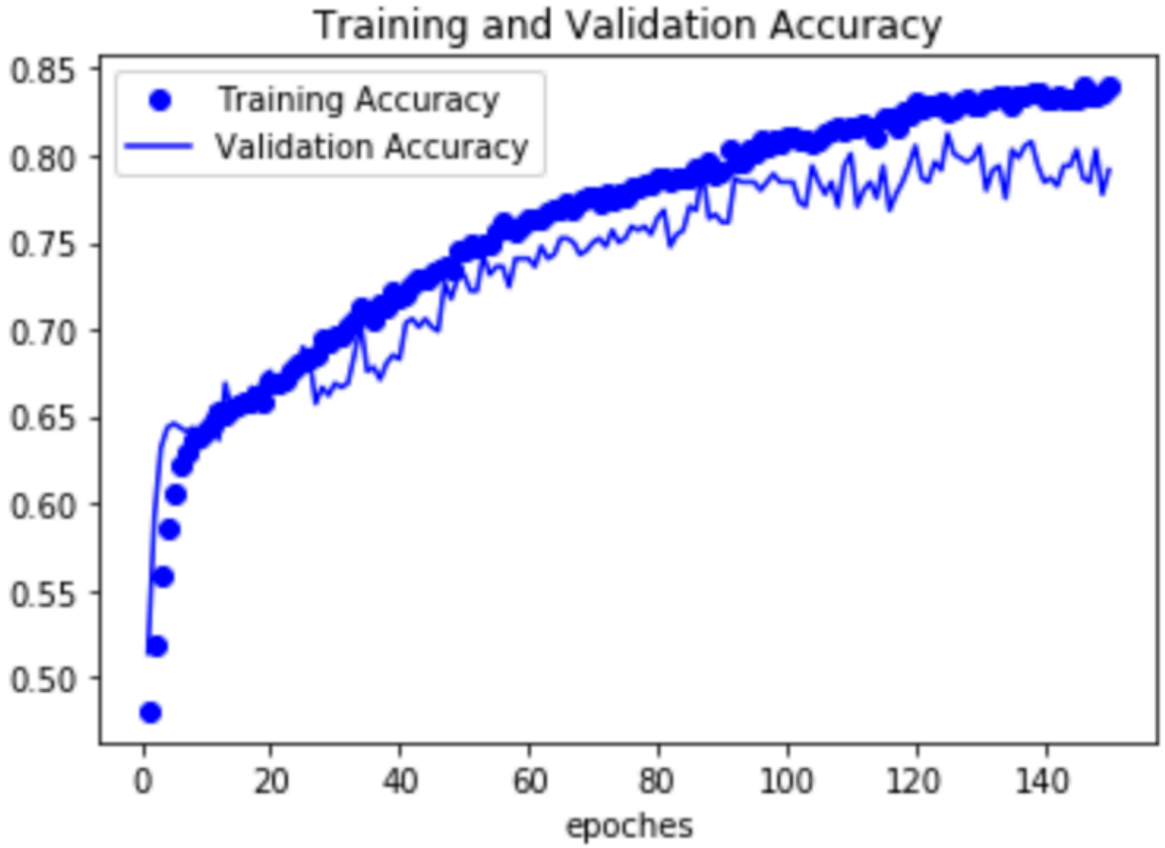


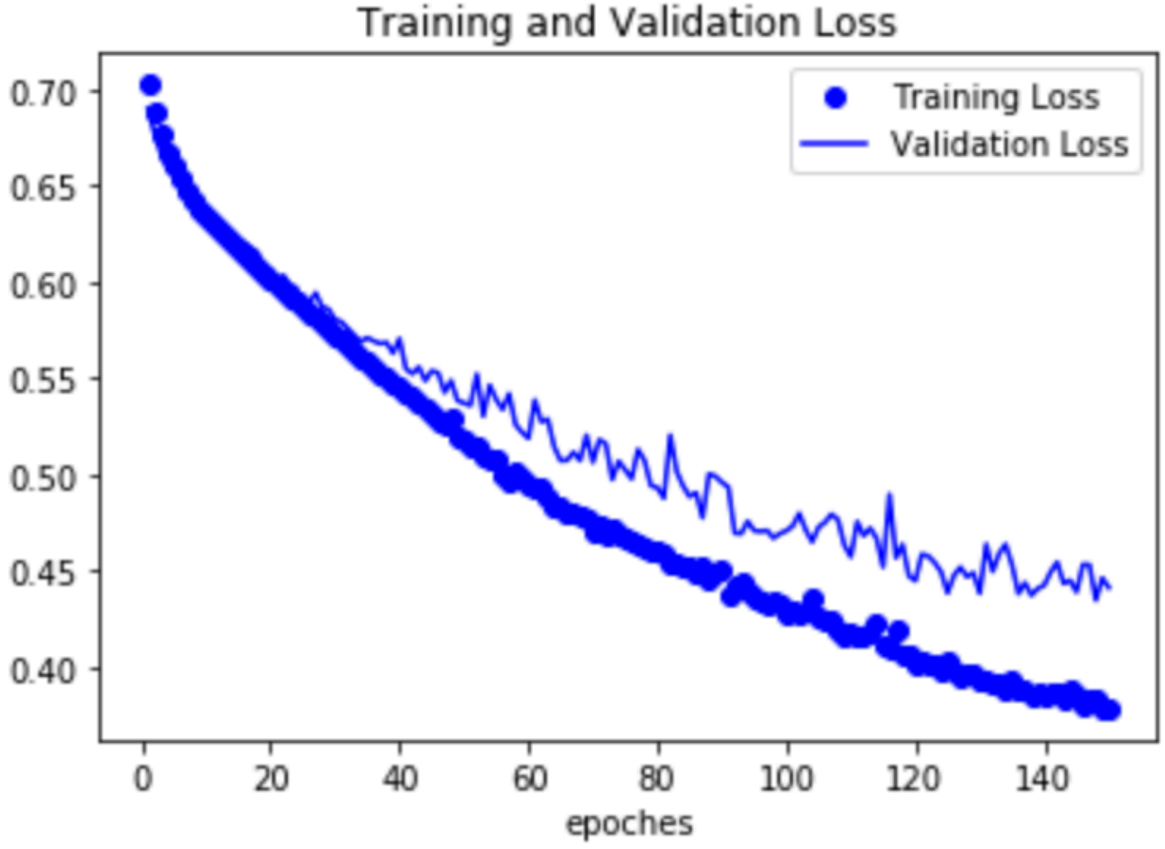


Only TFBS

[04_shuffling_on_padding_at_the_end_with_size_1000_bidirectional_only_TFBS.ipynb]

Accuracy 79%

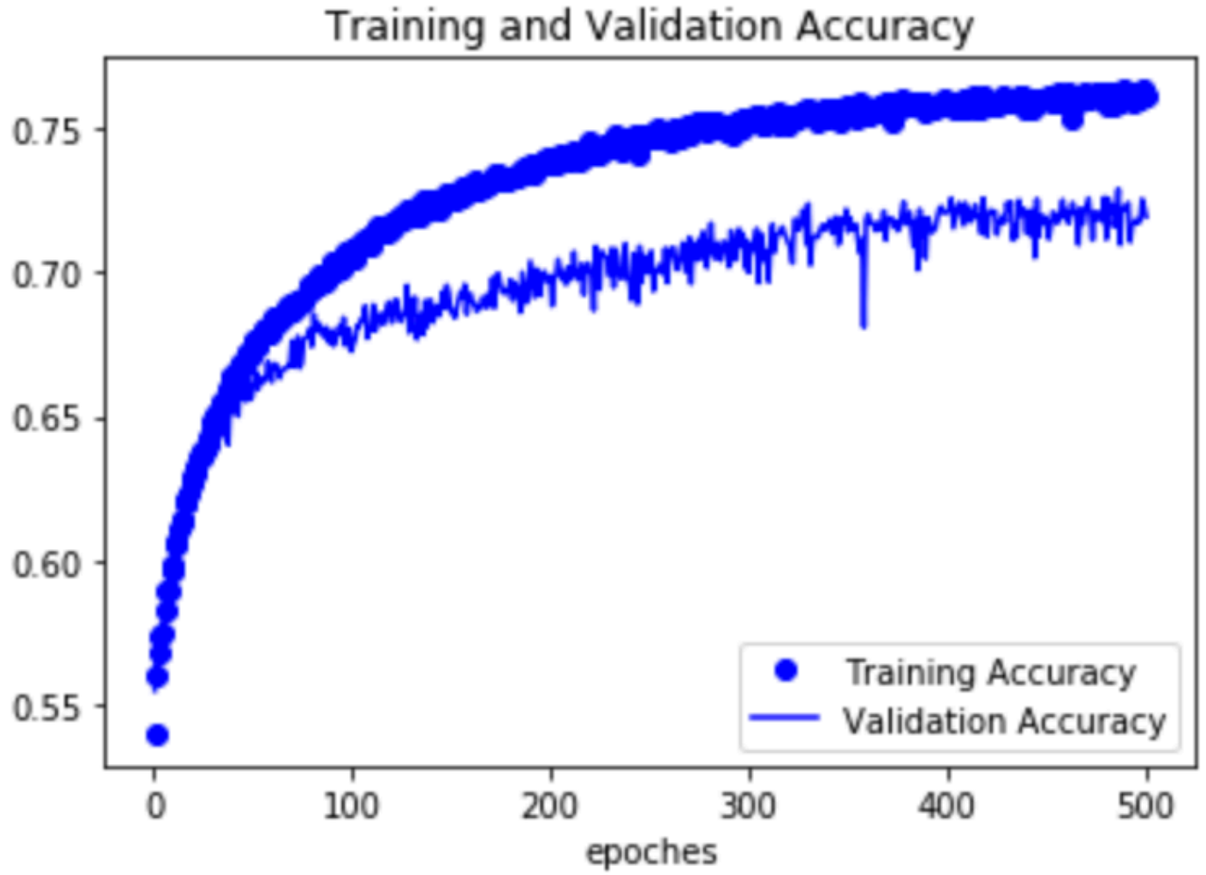


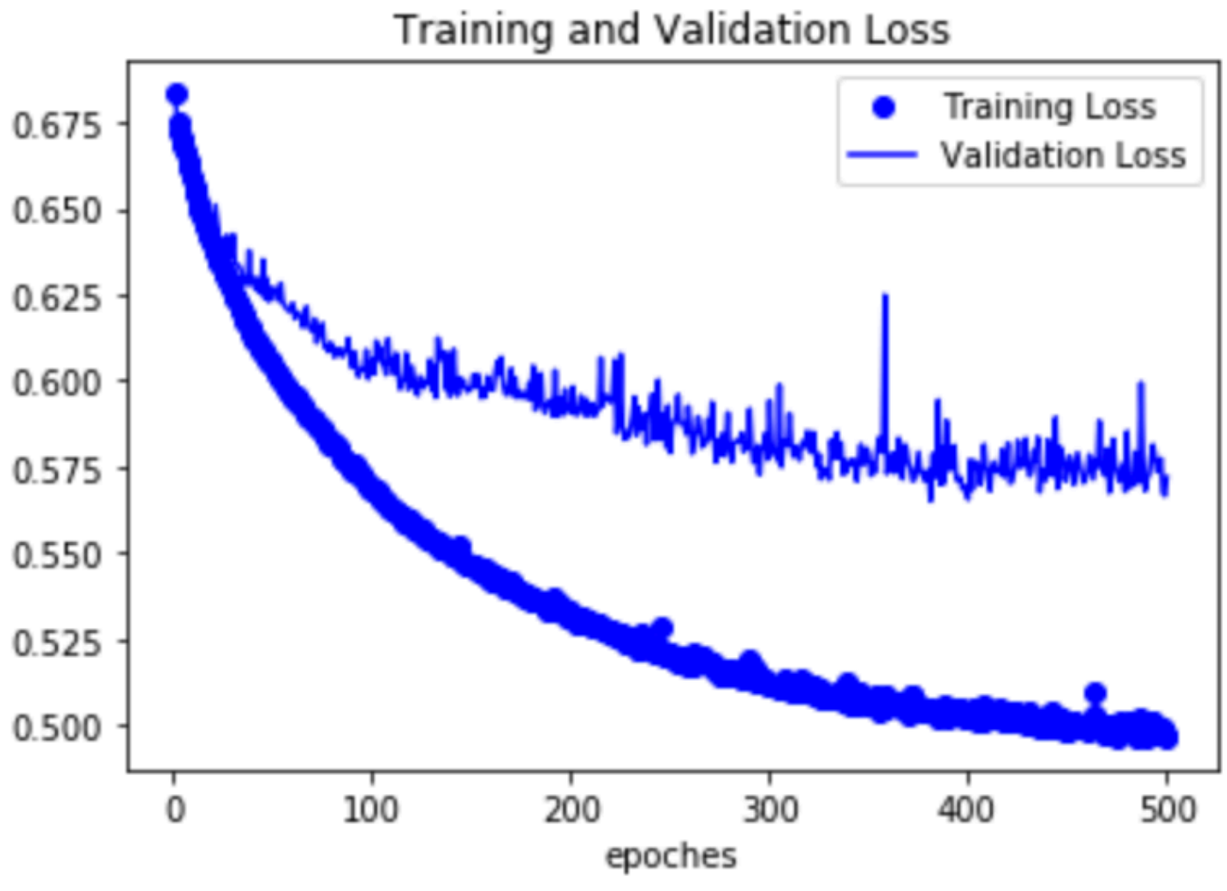


On whole dataset

[2019-02-10_whole_dataset_500_epochs.ipynb]

Accuracy 72%,
but kept climbing





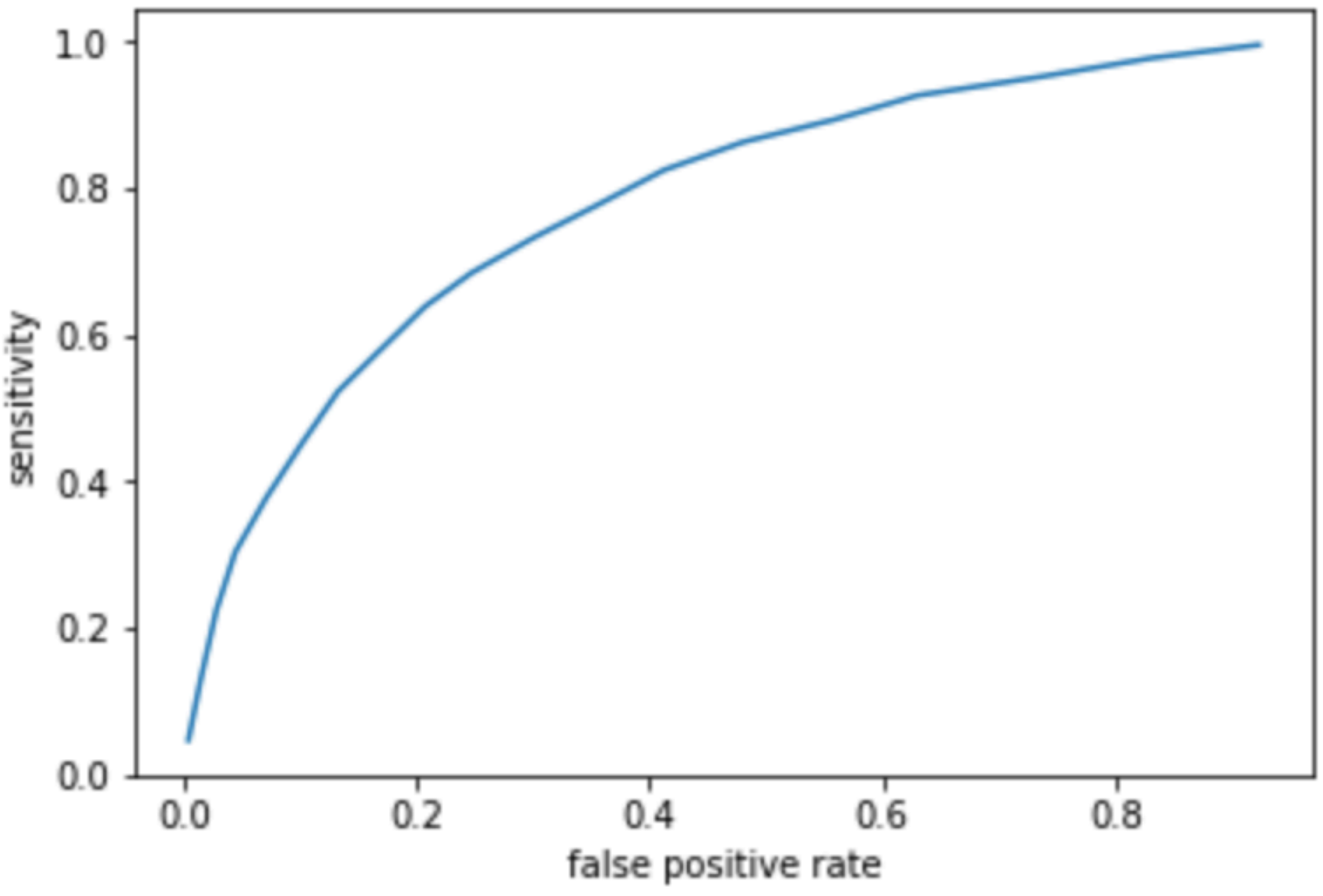
Further Evaluation

FDR

```
The false discovery rate is: 0.30776173285198555
```

(30% guessing a False positive)

C-Statistic




```
c-statistic = 0.7120263907710906
```

Agenda

I. Summary of Data Formatting
II. Bidirectional Experiments
III. Controls
    A. Random Sequences
    B. Random PWM