Deep Learning in Computational Biology

Final Project

Predicting PBM binding from HT-SELEX data

Roy Hirsch
Amit Zeligman

August 6, 2018

Main Goal:

Given a set of HT_SELEX data of different TF's. Learn a binding model for each TF and use it to rank PBM models.

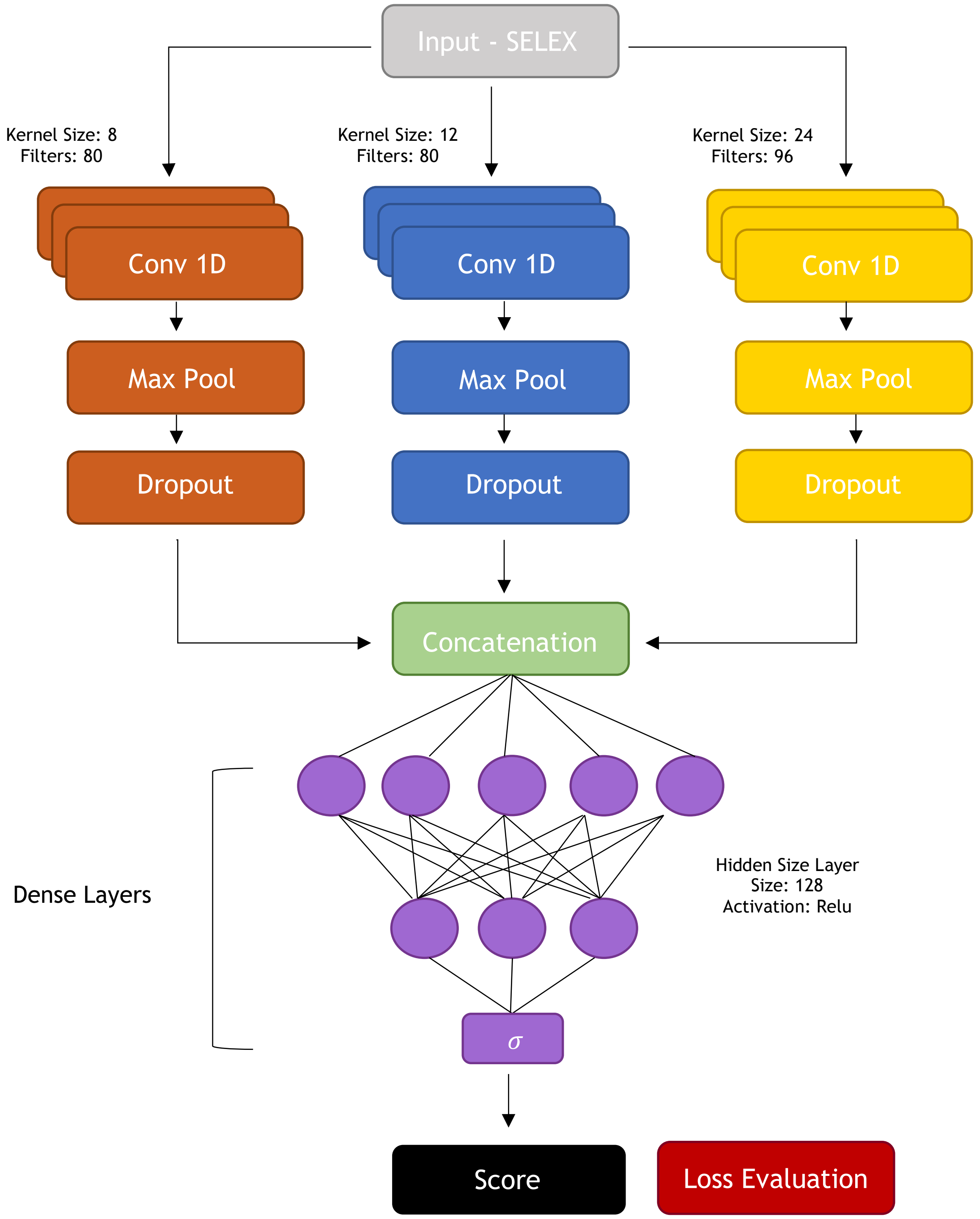
Input

- 123 training sets of different TF's. Each set include:
  - HT_SELEX: 4-6 sequence file with number of counts of each sequence.
  - Sorted PBM probes: a sorted List of sequences with their binding probs.
- 123 test sets (HT-SELEX data + unsorted PBM file).

Output

- A sorted PBM file – same sequences as in the input, only sorted.

Network Architecture[1]:



Parameters:

Learning Rate	1e-6
Learning Rate Decay	1e-6
Batch Size	32
Epochs	10
Conv Kernel Size	[8,12,24]
Conv Filter Size	[80,80,96]
Conv Strides	1
Pool Size	10
Conv Activation	Relu
Optimizer	Adam
Loss Function	Binary Cross entropy
Dropout	0.5
FC Hidden Size	128
Hidden Layer Activation	Relu
FC activation	Sigmoid

Parameter Search

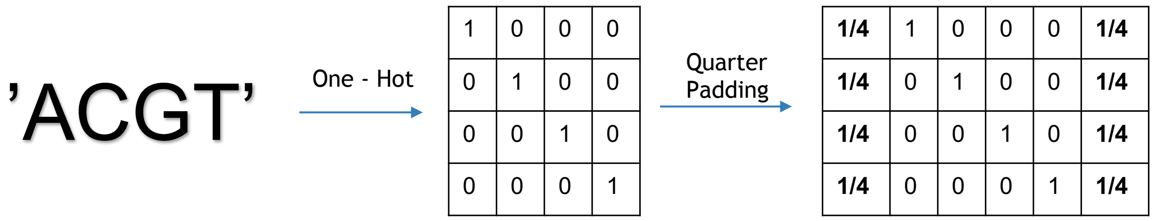
Depth	Dropout	FC Hidden Size	Lr decay	Max Pool	Optimizer	Samples Number	Mean AUPR	Std AUPR
[80, 80, 96]	**0.5**	**128**	**1.00E-06**	**10**	**adam**	**15**	**0.1389**	**0.1334**
[80, 80, 96]	0.5	128	1.00E-06	14	adam	20	0.1206	0.1373
[80, 80, 96]	0.5	128	1.00E-06	2	adam	15	0.1069	0.0977
[80, 80, 96]	0.5	256	1.00E-06	6	adam	15	0.1065	0.1087
[80, 80, 96]	0.5	128	1.00E-06	12	adam	20	0.1051	0.1074
[80, 80, 96]	0.5	128	1.00E-07	6	adam	15	0.1039	0.1045
[80, 80, 96]	0.5	256	1.00E-06	12	adam	20	0.1032	0.0946
[80, 80, 96]	0.25	32	0	4	adam	15	0.0883	0.0886
[80, 80, 96]	0.5	128	1.00E-06	6	adam	20	0.0857	0.0957
[80, 80, 96]	0.5	64	0	2	ada_delta	15	0.082	0.0674
[80, 80, 96]	0.5	32	1.00E-06	2	adam	15	0.0813	0.089
[80, 80, 96]	0	64	0	6	adam	15	0.0798	0.0694
[40, 40, 48]	0.25	64	0	2	adam	15	0.0781	0.0885
[80, 80, 96]	0.5	64	1.00E-06	6	adam	20	0.0767	0.0899
[40, 40, 48]	0.5	64	0	2	adam	15	0.0758	0.0722
[80, 80, 96]	0	32	0	2	adam	15	0.0757	0.0673
[80, 80, 96]	0.5	128	0	6	ada_delta	15	0.0746	0.0681
[40, 40, 48]	0	32	1.00E-07	6	adam	15	0.0724	0.0544
[80, 80, 96]	0.25	64	1.00E-06	4	ada_delta	15	0.0636	0.0531
[40, 40, 48]	0	32	0	6	ada_delta	15	0.0554	0.0507
[80, 80, 96]	0.25	32	1.00E-07	2	ada_delta	15	0.0552	0.0497
[40, 40, 48]	0	64	1.00E-06	4	ada_delta	15	0.0529	0.0407

Preprocessing

- One – Hot encoding + quarter padding:

The sequences encoded two one-hot format and then padded with quarters such that:

$$Input\ size = [\max(PBM\ length,\ SELEX\ length),\ 4]$$



’ACGT’

One - Hot

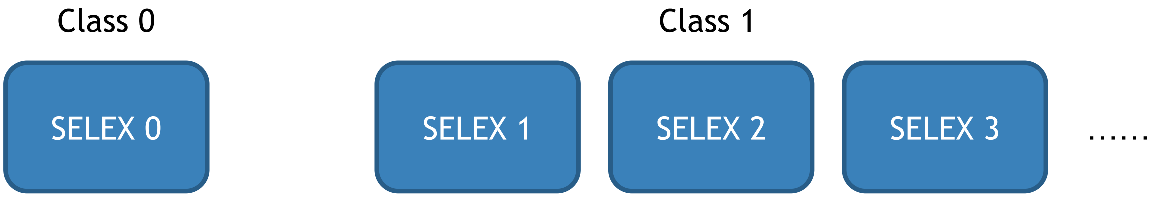
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Quarter Padding

1/4	1	0	0	0	**1/4**
1/4	0	1	0	0	**1/4**
1/4	0	0	1	0	**1/4**
1/4	0	0	0	1	**1/4**

- Labeling + Data Size Balance:

The training data were divided into two classes:
Class 0 – SELEX 0, Class 1 – all the other SELEX.



Class 0

Class 1

SELEX 0

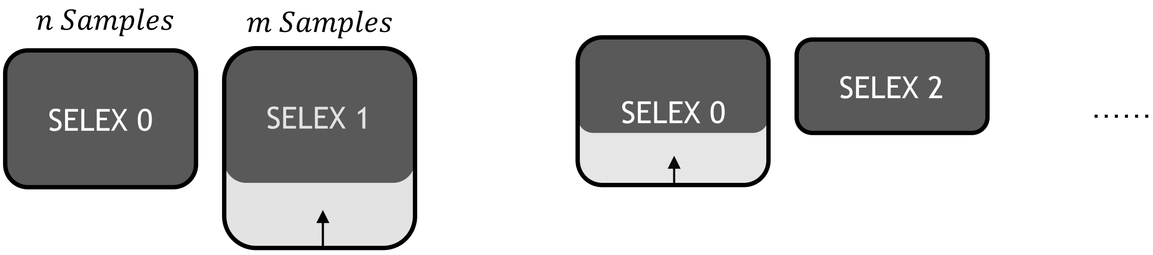
SELEX 1

SELEX 2

SELEX 3

……

Each SELEX pair (class 0 + class 1) filtered such that the number of samples for class 0 and class 1 will be equal.



$n\ Samples$

$m\ Samples$

SELEX 0

SELEX 1

SELEX 0

SELEX 2

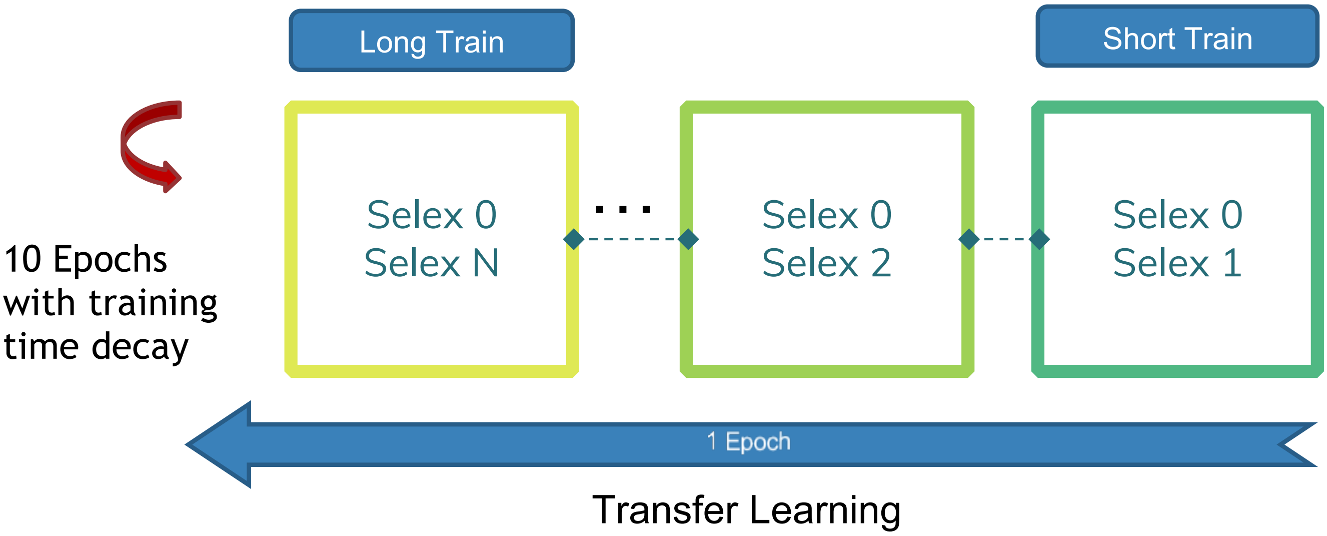
……

$$Number\ of\ Sample[i] = \min(n, m)$$
$$i = 1 \ldots Number\ of\ SELEX$$

Training Scheme

The model was trained on each SELEX pair individually (i.e. loss evaluation and backpropagation evaluated after each SELEX pair train). In order to let the model to learn “strong” features, short training time (training steps) was given to the first SELEX pair (0+1), while long time given to the last pair.

The model was trained for 10 Epochs. Each epoch composed from N-1 training on SELEX pairs, Where N is the number of SELEX cycles.
The training time(training steps) decreased linearly after each epoch.



Training Results

AUPR:

Average	High	Low	Std
0.0753	0.5819	0.0021	0.0072

Performance

Average memory usage:5.393 GB
Average run-time: 1h 7m (01:07:00)
Average CPU: 2367%

References

[1] High-Order Convolutional Neural Network Architecture for Predicting DNA-Protein Binding Sites, QinHu Zhang, Lin Zhu, and De-Shuang Huang, Senior Member, IEEE

Appendix

Detail Network weights and structure:

```
Layer (type)                  Output Shape        Param #          Connected to
==================================================================================================
input_1 (InputLayer)          (None, 36, 4)       0
__________________________________________________________________________________________________
conv1d_1 (Conv1D)             (None, 36, 80)      2640             input_1[0][0]
__________________________________________________________________________________________________
conv1d_2 (Conv1D)             (None, 36, 80)      3920             input_1[0][0]
__________________________________________________________________________________________________
conv1d_3 (Conv1D)             (None, 36, 96)      9312             input_1[0][0]
__________________________________________________________________________________________________
max_pooling1d_1 (MaxPooling1D)  (None, 3, 80)       0            conv1d_1[0][0]
__________________________________________________________________________________________________
max_pooling1d_2 (MaxPooling1D)  (None, 3, 80)       0            conv1d_2[0][0]
__________________________________________________________________________________________________
max_pooling1d_3 (MaxPooling1D)  (None, 3, 96)       0            conv1d_3[0][0]
__________________________________________________________________________________________________
dropout_1 (Dropout)           (None, 3, 80)       0          max_pooling1d_1[0][0]
__________________________________________________________________________________________________
dropout_2 (Dropout)           (None, 3, 80)       0          max_pooling1d_2[0][0]
__________________________________________________________________________________________________
dropout_3 (Dropout)           (None, 3, 96)       0          max_pooling1d_3[0][0]
__________________________________________________________________________________________________
concatenate_1 (Concatenate)     (None, 3, 256)      0          dropout_1[0][0]
                                                               dropout_2[0][0]
                                                               dropout_3[0][0]
__________________________________________________________________________________________________
flatten_1 (Flatten)           (None, 768)         0              concatenate_1[0][0]
__________________________________________________________________________________________________
dense_1 (Dense)               (None, 128)         98432          flatten_1[0][0]
__________________________________________________________________________________________________
dropout_4 (Dropout)           (None, 128)         0              dense_1[0][0]
__________________________________________________________________________________________________
dense_2 (Dense)               (None, 1)           129            dropout_4[0][0]
==================================================================================================
Total params: 114,433
Trainable params: 114,433
Non-trainable params: 0
__________________________________
```