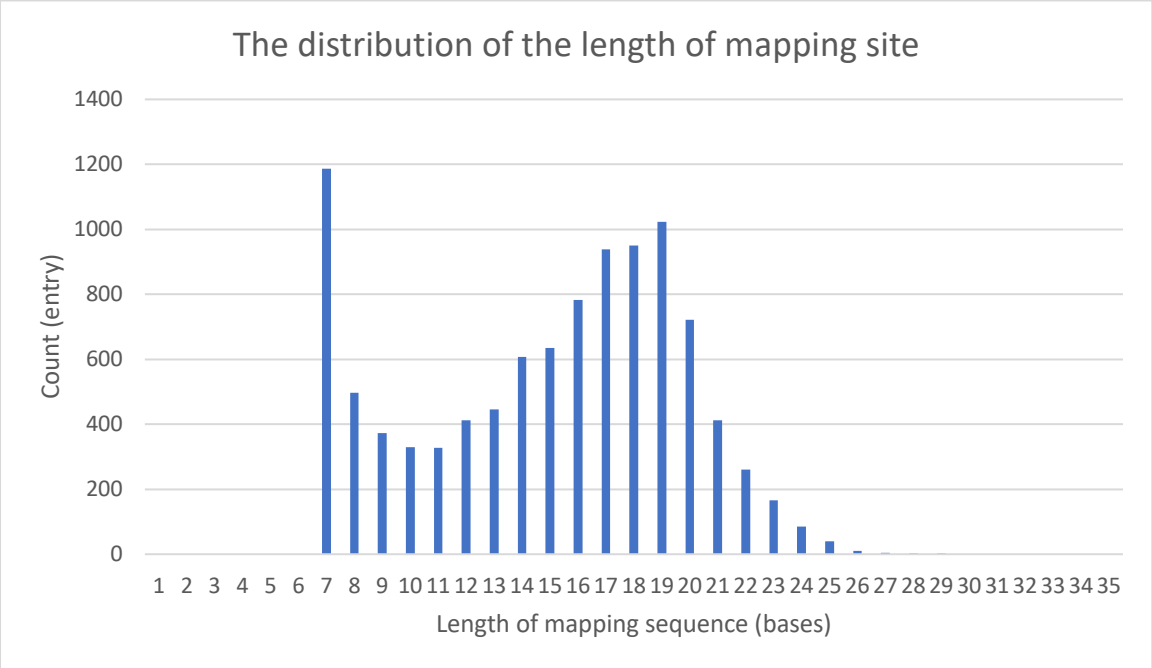
Term Project Report (1)

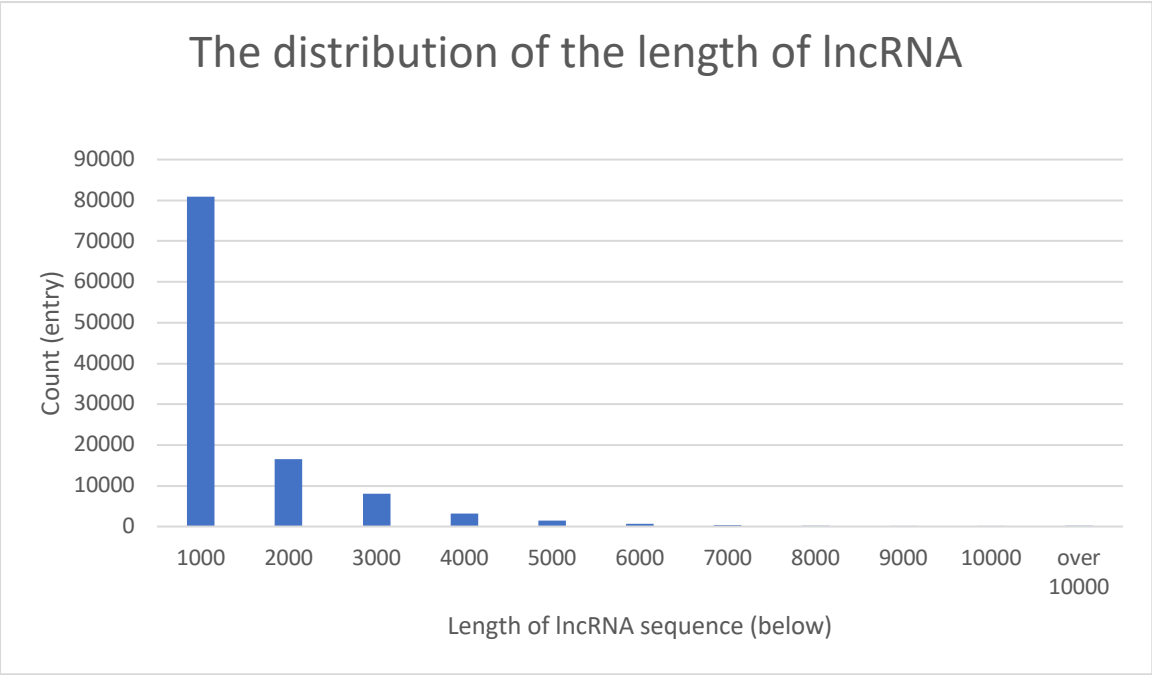
Data

Using data from starBase v2.0 and lncipedia v3.1 to create dataset of mapping site in long non-coding RNA.

Example

AGAGGTGCCTCCCTTCCTTGAAACTTCTTCACCTTTGCTTCAAGAGCACCACAGAGCAGATTC
TTCAAGGCAGAGATATACAAGAATCCCTGTGGAAAACCTCATGAACTCAAGGTCATGCAGAGACTTAGG
GAAGAACTGGGACTGGCTTATCAAGACTCTACACTACTGACCCAGGGCTTGGCAGTTTCTGTCCTAGAAA
TCAGAACACGCAGCTCATGCATCACCGAGAAGGTGGTCAGCCTAAGAAGTGGTTCCCCACTCTTTGTTCT
AAGATATTCCTCTCTCAGA





Preparing Dataset

I want the model to find the connection between each bases in every possible of sequence that can occur to do that the dataset consist of 4 type of data entry.

- Backward
- Forward
- Cover
- Negative Sampling

Note: Example of cutting size equal to 20

Raw-sequence

AGAGGTGCCTCCCTTCCTTGAAACTTCTTCACCTTTGCTTCAAGAGCACCACAGAGCAGATTC
TTCAAGGCAGAGATATACAAGAATCCCTGTGGAAAACCTCATGAACTCAAGGTCATGCAGAGACTTAGG
GAAGAACTGGGACTGGCTTATCAAGACTCTACACTACTGACCCAGGGCTTGGCAGTTTCTGTCCTAGAAA
TCAGAACACGCAGCTCATGCATCACCGAGAAGGTGGTCAGCCTAAGAAGTGGTTCCCCACTCTTTGTTCT
AAGATATTCCTCTCTCAGA

Backward

CAGGGCTTGGCAGTTTCTGT

CCAGGGCTTGGCAGTTTCTG

CCCAGGGCTTGGCAGTTTCT

ACCCAGGGCTTGGCAGTTTC

GACCCAGGGCTTGGCAGTTT

TGACCCAGGGCTTGGCAGTT

CTGACCCAGGGCTTGGCAGT

ACTGACCCAGGGCTTGGCAG

TACTGACCCAGGGCTTGGCA

CTACTGACCCAGGGCTTGGC

ACTACTGACCCAGGGCTTGG

TACTACTGACCCAGGGCTTG

CTACTACTGACCCAGGGCTT

ACTACTACTGACCCAGGGCT

Forward

TTGGCAGTTTCTGTCCTAGA

TGGCAGTTTCTGTCCTAGAA

GGCAGTTTCTGTCCTAGAAA

GCAGTTTCTGTCCTAGAAAT

CAGTTTCTGTCCTAGAAATC

AGTTTCTGTCCTAGAAATCA

GTTTCTGTCCTAGAAATCAG

TTTCTGTCCTAGAAATCAGA

TTCTGTCCTAGAAATCAGAA

TCTGTCCTAGAAATCAGAAC

CTGTCCTAGAAATCAGAACA

TGTCCTAGAAATCAGAACAC

GTCCTAGAAATCAGAACACG

TCCTAGAAATCAGAACACGC

Cover

CTTGGCAGTTTCTGTCCTAG

GCTTGGCAGTTTCTGTCCTA

GGCTTGGCAGTTTCTGTCCT

GGGCTTGGCAGTTTCTGTCC

AGGGCTTGGCAGTTTCTGTC

Negative Sampling

CTACACTACTGACCCAGGGC

TCTACACTACTGACCCAGGG

CTCTACACTACTGACCCAGG

ACTCTACACTACTGACCCAG

GACTCTACACTACTGACCCA

.....

Methods and models

Using deep learning technique to find the connection between each bases all we need is a good datasets and the proper neural network model.

The model

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 500, 50)           250
_________________________________________________________________
dropout_1 (Dropout)          (None, 500, 50)           0
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 491, 600)          300600
_________________________________________________________________
dropout_2 (Dropout)          (None, 491, 600)          0
_________________________________________________________________
bidirectional_1 (Bidirection (None, 491, 600)          1621800
_________________________________________________________________
dropout_3 (Dropout)          (None, 491, 600)          0
_________________________________________________________________
bidirectional_2 (Bidirection (None, 600)               1621800
_________________________________________________________________
dropout_4 (Dropout)          (None, 600)               0
_________________________________________________________________
dense_1 (Dense)              (None, 500)               300500
=================================================================
Total params: 3,844,950
Trainable params: 3,844,950
Non-trainable params: 0
```

1st Layer: Embedding layer with window size equal to 500 bases and number of dimension is 50

Reason: I also want the model to find the relationship between A, T, C, G and N bases incase there are a hidden information we don't know about it now.

2nd Layer: Dropout Layer

Reason: I want to avoid overfitting of the model because I use a lot of neural units.

3rd Layer: 1-Dimension Convolutional Layer with 600 filter and kernal size equal to 10 bases

Reason: because the length of mapping size is arround 10-20 bases. (show in the 1st Chart)

4th Layer: Dropout Layer

5th Layer: Bi-directional-GRU with 300 neural units (forward) and 300 neural units (backward)

Reason: I choose bi-directional RNN because I assume that the connection between each bases (A, T, C, G, N) can be both forward and backward direction.

6th Layer: Dropout Layer

7th Layer: Bi-directional-GRU with total 600 unit (backward and forward)

8th Layer: Dropout Layer

9th Layer: Output Layer with 500 neural unit (equal to input layer size)

Reason: because I want the model to predict that in each position of a given sequence it's a mapping size or not.

Using the model

Before using the model we need to cut the sequence of lncRNA into a 500 length long of sequence and then feed it into our model. There are so many method that can be done but cutting the sequence in non-overlapping manner is recommended.

Current Results

Old result (when using window size equal to 50)

Macro Accuracy is 0.92

Micro F1 score of '1'

Correct (1): 1014950

Correct (0): 8358868

Wrong (0->1): 288291

Miss (1->0): 412541

Recall: 0.7110027313657319

Precision: 0.7787891878785275

F-score: 0.7433537967109186

New result (when using window size equal to 500)

Macro Accuracy is 0.98

P.S. the model is in development

Ongoing process

There are a lot of hyperparameter that I want to try first "input window size" which I think it's the most important parameter because I don't really know what is the connection between each ATCGN and how much the information is need to help the model figure it out but more information is given (long sequence) more time and computation is needed.