Application of Deep Learning Models to MicroRNA Transcription Start Site Identification

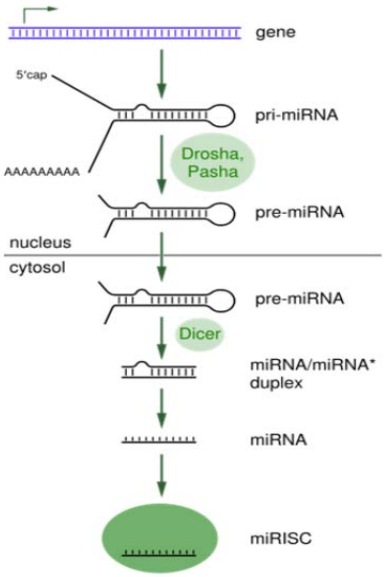
Clayton Barham, Mingyu Cha, Xiaoman Li, Haiyan Hu

microRNA

- Small non-coding RNA, ~22 nucleotides, abbreviated as miRNA
- Involved in almost all key biological processes, such as development, differentiation, and so on.
- RNA silencing and post-transcriptional regulation of gene expression via base-pairing with complementary sequences within mRNA of protein coding genes

Crucial steps in miRNA processing

- Transcribed by RNA polymerases II or III as pri-miRNA (primary miRNA)
- Processed by RNase II enzyme Drosha to miRNA precursor (pre-miRNA)
- Transported to cytoplasm, and RNase II enzyme Dicer cleaves off the double strand of the hairpin to form mature miRNA
- Forms miRNPs (miRNA-protein complexes) and binds to partially complementary sites



https://www.lcsciences.com/discovery/a-key-post-transcriptional-modification-promotes-the-initiation-of-mirna-biogenesis/

miRNA TSS prediction Hard

- Difficulty of predicting promoters from short conserved sequence
- features without producing a high number of false positives
- Pri-miRNAs are several kilobases long, and rapidly cleaved in the nucleus by the enzyme Drosha
- Limited experimental detection and annotation of miRNA promoter
- Recent studies indicate that intronic miRNA are not necessarily cotranscribed with host genes

Problem Description

- Predict miRNA TSS
- Predict features from raw sequence data

Dataset Description

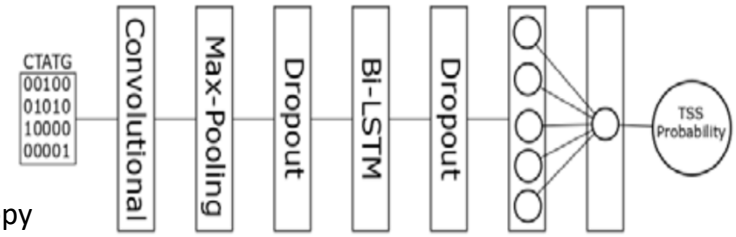
- Compilation of 7,610 CAGE TSS peak regions
  - Width smaller than 10bps
  - At least 500bp away from each other
  - Have PolII, Dnase and H3K4m3 active gene markers
- Positive training data: the central 100bp of the compiled TSS regions
- Negative training data:
  - 100bp flanking of the positive training regions
  - 100bp randomly sampled from intergenic regions
- Convert sequences to 100x4 one hot encoding
  - A C T G G C T A A C

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

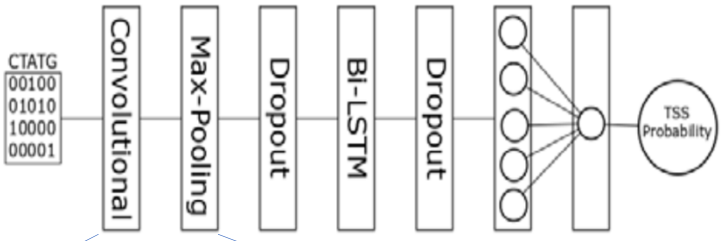
Method Description –LSTM Model Architectures

- LSTM (long short term memory)-based Architecture:
  - Input
  - Convolutional Layer
  - Max Pooling Layer
  - Dropout Layer
  - Bi-LSTM layer
  - Dropout Layer
  - Dense Layer
  - Sigmoid Output
- Loss function: binary cross-entropy
- Activation function: Rectified Linear Unit function

- Input: 100*4 one-hot matrix encoding the 100bp sequence
- Output: a probability of the input containing a TSS



Method Description – Convolution & Max Pooling

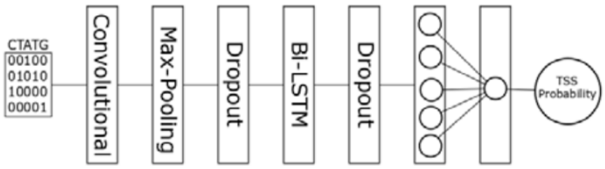


- Convolutional Layer:
  - Neural network layer using weight sharing
  - Scan input with a number of kernels
  - Recognizes motifs in DNA sequence input

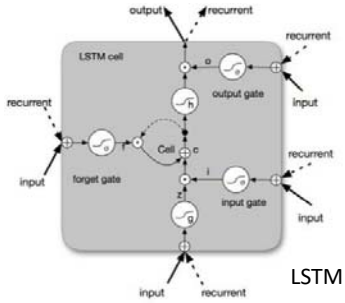
- Max Pooling Layer:
  - Reduces size of input
  - Out of a set of pool_size elements, select the max

Method Description – Bi-LSTM

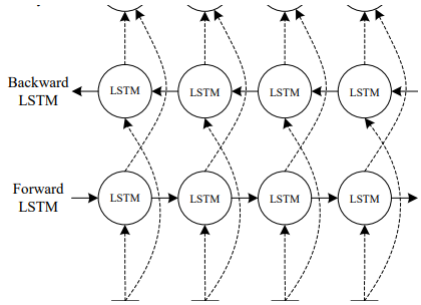
- LSTM:
  - Long-Short Term Memory
  - Bi-LSTM = Bidirectional LSTM
  - Learns 'regulatory grammar' – spatial patterns of motifs
    - Use this to refine probability of a given motif at a given position



Recognizes long-term relationships between motifs => regulatory grammer



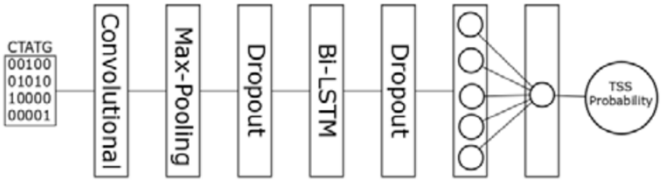
LSTM unit



Bi-LSTM

arXiv:1603.01354

Method Description – Dense & Sigmoid

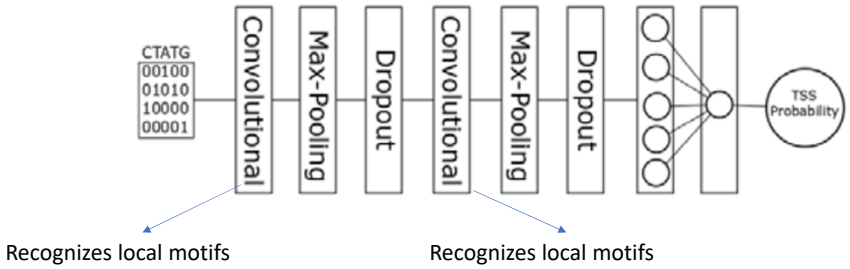


- Dense Layer:
  - Standard neural network layer
- Sigmoid Layer:
  - Dense layer with sigmoid activation function
  - Convert feature activation values to probabilities

Method Description – CNN Model Architectures

- CNN (convolutional neural network)-based Architecture:
  - Input
  - Convolutional Layer
  - Max Pooling Layer
  - Dropout Layer
  - Convolutional Layer
  - Max Pooling Layer
  - Dropout Layer
  - Dense Layer
  - Sigmoid Output

- Input: 100*4 one-hot matrix encoding the 100bp sequence
- Output: a probability of the input containing a TSS



Results: Flanking Data Set

Data Set	Precision	Recall	F1-Score	Accuracy
LSTM 0	0.8562	0.3671	0.5139	0.7749
LSTM 1	0.3529	0.1094	0.1670	0.6545
LSTM 2	0.9280	0.2984	0.4516	0.7358
LSTM 3	0.9437	0.2694	0.4192	0.7527
LSTM 4	0.9810	0.2640	0.4161	0.7420
Average	0.8124	0.2617	0.3959	0.7320
Average w/o LSTM 1	**0.9272**	0.2998	0.4530	**0.7513**
CNN 0	0.9302	0.3288	0.4858	0.7744
CNN 1	0.9522	0.2804	0.4332	0.7456
CNN 2	0.8725	0.3167	0.4647	0.7340
CNN 3	0.8649	0.3003	0.4458	0.7527
CNN 4	0.9809	0.2615	0.4129	0.7411
Average	0.9201	0.2975	0.4497	0.7496
Average w/o CNN 1	0.9121	**0.3018**	**0.4535**	0.7506
SVM 0	0.2283	0.2342	0.2312	0.4951
SVM 1	0.2931	0.2830	0.2879	0.5147
SVM 2	0.2969	0.2716	0.2837	0.5
SVM 3	0.2322	0.2493	0.2405	0.4782
SVM 4	0.2680	0.2423	0.2545	0.5058
Average	0.2637	0.2561	0.2598	0.4988
Average w/o SVM 1	0.2564	0.2494	0.2528	0.4948

- Test on flanking data set shows deep learning models are able to distinguish TSS regions from their neighboring regions.
- Comparisons between LSTM, CNN, SVM show deep learning models have much better performance.

Results: Intergenic Data Set

Data Set	Precision	Recall	F1-Score	Accuracy
LSTM 0	0.972	0.3408	0.5047	0.7882
LSTM 1	0.7146	0.3981	0.5113	0.7513
LSTM 2	0.9007	0.3267	0.4795	0.7638
LSTM 3	0.9390	0.3072	0.4629	0.7620
LSTM 4	0.9289	0.2776	0.4274	0.7513
Average	**0.8911**	0.3301	0.4817	0.7633
CNN 0	0.9713	0.3324	0.4953	0.7855
CNN 1	0.9081	0.3356	0.4901	0.7718
CNN 2	0.8974	0.3267	0.4790	0.7633
CNN 3	0.8076	0.3684	0.5059	0.7598
CNN 4	0.8601	0.3347	0.4818	0.7593
Average	0.8889	**0.3395**	**0.4913**	**0.7679**
SVM 0	0.2297	0.2496	0.2392	0.4973
SVM 1	0.2154	0.2201	0.2177	0.4831
SVM 2	0.2344	0.2253	0.2298	0.4969
SVM 3	0.2304	0.234	0.2322	0.4831
SVM 4	0.2347	0.2297	0.2322	0.4920
Average	0.2289	0.2312	0.2303	0.4905

- Test on intergenic data set shows deep learning models are able to distinguish TSS regions from intergenic regions.
- Comparisons between LSTM, CNN, SVM show deep learning models again have much better performance.

Results: Cell-line Specific Predictions

	GM12878	HeLa-S3	HepG2	K562
LSTM	623 / 57	757 / 64	637 / 65	718 / 71
CNN	616 / 57	764 / 58	628 / 59	719 / 73

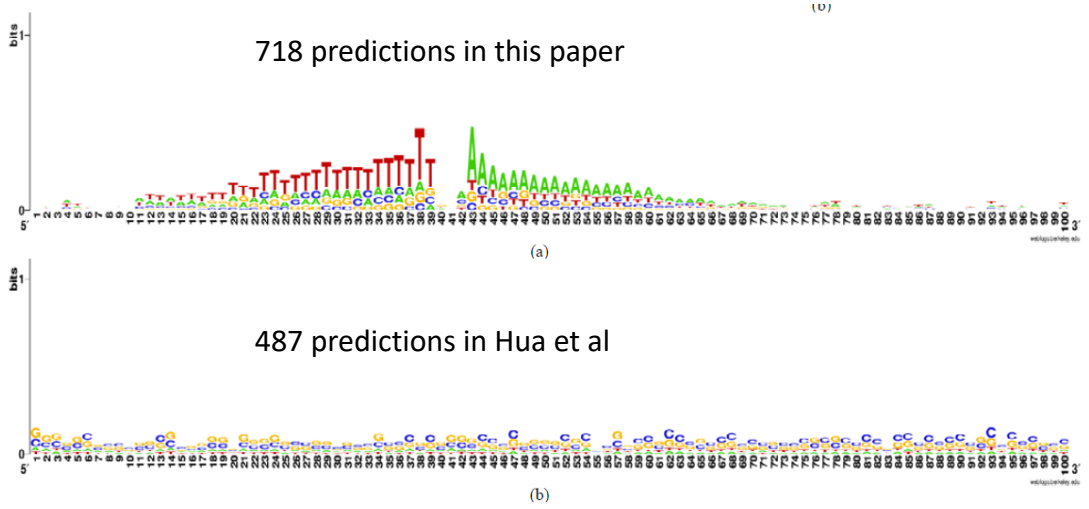
ofTRUE / # of FALSE predictions

	GM12878	HeLa-S3	HepG2	K562
LSTM	91.62%	92.20%	90.74%	91.00%
CNN	91.53%	92.94%	91.41%	90.78%

Accuracy of predicted results by cell lines

- We separate the test data by cell lines to compare results in a cell-line specific manner.
- Test on flanking data set shows deep learning models are able to identify cell-line relevant TSS regions

Results: Compare with Hua et al: the sequence feature (TATA) identification from TSS predictions in K562 cell line



https://weblogo.berkeley.edu/logo.cgi

Conclusions

- Both LSTM and CNN models provided high-resolution miRNA TSS predictions for testing dataset obtained from FANTOM project.
- The trainings partition for each data set consisted of approximately 18,000 training samples, which is a small training set by the standards of typical deep learning training sets. More positive examples for either network will further improve its ability to identify TSSs.
- We focused on only narrow peaks of CAGE TSS regions. Our model indicate any true positive sample which lacked a sufficiently clearly defined TATA box in the expected position would not have been recognized as a positive sample. This suggests a potential avenue for improving the performance of the test models by encouraging them to consider additional features.

Questions?