Comparison of Lin28a binding site prediction models : HMM and CNN

Introduction

Lin28a is a conserved RNA binding protein, identified from *C. elegans* to mammalian animals, including mouse and human. In the past studies, Lin28a was considered as a protein that mainly acts as a suppressor of let-7 miRNA biogenesis. Recently, J. Cho *et al*.(2012) suggested that Lin28a has additional function as suppressor of ER-Associated translation in stem cells.

Since Lin28a is associated with the Yamanaka factors, known as the key genes in induced Pluripotent Stem Cell(iPSC) generation, predicting the Lin28a binding site may provide future insight and potential key genes in cell differentiation. So for the task of predicting potential Lin28a-binding-mRNAs, two prediction models, Hidden Markov Model(HMM) and CNN(Convolutional Neural Network), were applied and compared.

Dataset

For dataset generation, first all the protein-coding genes existing in (+) strands were extracted using samtools from gencode file. Then, after removing mitochondrial genes, rest of the genes were used for CLIP-seq data pileup with samtools.

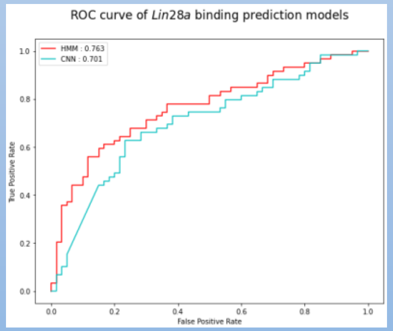
After calculating Shannon's entropy of each gene positions, cut-off of 1.2 was used for locating Lin28a-binded regions. Among 5,000 genes, only 129 showed positions with over 1.2 entropy, and the 18-mers of -8~+9 region were used as binding site answers.

The same amount of randomly-generated sequences were used as negative dataset. Total dataset was split into train:test data in portion of 7:3.

Results

The AUROC curve of binding prediction models are shown in Fig 3. The HMM(red) model shows better performance over CNN(cyan) in binding prediction, with 0.763 AUROC over 0.701 of CNN's.

It seems that relatively small number of sequences (~200s) for training and short length(18-mer) limited CNN's representation learning contrast to HMM's probabilistic approach.



Prediction Models

Hidden Markov Model

Hidden Markov Model is as probabilistic model that learns and predicts the possible hidden state sequence through Baum-Welch algorithm, which is a special type of expectation-maximization structure.

The model structure and initial parameters of trained HMM is shown in Fig 1. Classification is done using logistic regression model with HMM forward probability as feature.

The HMM model was generated and trained using python package hmmlearn version 0.2.5.

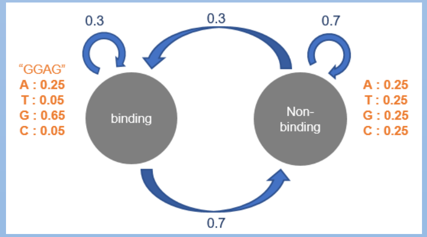


Fig 1. HMM model structure. Gray nodes indicate the two states, blue arrows represent transition probabilities and orange numbers are emission probabilities for each nucleotide.

Convolutional Neural Network

Convolutional Neural Network is a convolution layer-based neural network, mainly used for dealing 2D-type data and learning special information to generate feature vectors from DNA sequences.

For this model, 1d-convolutional layer was adopted as the first layer, directly connected to the encoded input layer. Kernel of size 8 is given with stride of 2. Convolutional layer is followed by global max-pooling layer and one layer of 16 perceptrons, with one sigmoid-activated neuron in the end. The overall model structure is shown in Fig 2.

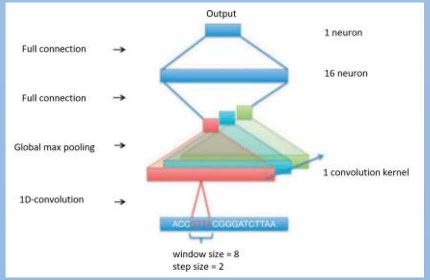


Fig 2. CNN model structure

Conclusion

Predicting which mRNAs bind to Lin28a may suggest potential key genes in cell differentiation. For this task, two models were used for comparison, HMM and CNN. HMM show characteristics of a probabilistic machine learning model, when CNN is a deep neural network is a non-linear model based on convolutional layer.

Results show that HMM-based prediction is more promising, indicating that on small(~200s) and short(18-mer) dataset, probabilistical model performs better than deep neural network.