A Convolutional Neural Network to Predict Histone Modification from DNA Sequence and Methylation Data

Huan-Kai Yang & Paul Horton

Institute of Computer Science and Information Engineering,
National Cheng Kung University

ww1212332@gmail.com

Abstract

Nowadays, many advanced measurement methods have been developed to observe the exact binding site of histone modifications. However, it is a very time-consuming and expensive experiment to measure different histone modifications on each cell line. In order to overcome this dilemma, recent researches have been many methods to predict various histone modifications through DNA sequences. However, DNA sequences cannot provide any cell line-specific information, which leads to a big bottleneck in predicting histone modifications in different cell lines. Therefore, this study introduced DNA sequences and methylation data, and utilized convolutional neural network (CNN) to improve the prediction of histone modifications in different cell lines. Based on the experimental results and analysis, we confirmed that our method is helpful for predicting histone modification in different cell lines.

Introduction

Histone modifications are mainly profiled by chromatin-immunoprecipitation followed by sequencing (ChIP-seq). This method uses corresponding antibody for histone modifications or specific DNA-binding proteins to identify enriched loci within genome. The advantage of ChIP-seq is the higher resolution and lower noise. But, the disadvantage of ChIP-seq is still expensive and time-consuming, because it requires lots of tissues, so that profiling rare biological samples is constrained.

Data imputation methods, that utilizing relation among different kinds of biological data to reconstruct missing value, are often used appropriately in this scenario. Previous researches have been many data imputation methods to predict histone modifications through DNA sequences. However, DNA sequences cannot provide any cell line-specific information, because the basic nucleotide sequence of the genomes of diverse cell types (e.g. information processing neurons, protective skin cells, etc.) is essentially the same. How then can these cell types be distinct? The answer is thought to largely lie in differences in so called "epigenetic marks", namely DNA methylation and various histone modifications.

Main Objectives

DNA methylation and histone modification not only are individually involved in several key biological processes but also have special interaction between themselves [1]. We hope that is able to utilize their special interaction to predict histone modifications, while DNA methylation data supplies cell-specific information. Therefore, the objectives of this researches are listed as follow:

- We will design a data imputation method based on deep-learning for histone modifications.
- Test whether introducing DNA methylation is helpful to predict in cell line-specific cases.

Materials and Methods

In our problem formulation, we define prediction of six histone modifications in a window as a binary classification task, given a fixed-length segment of DNA sequence and methylation signal. This is known as a multi-label classification task, where multiple labels can be positive at once.

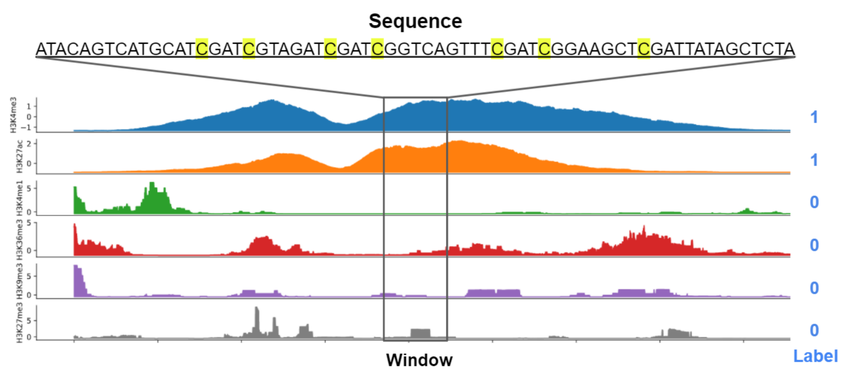

Figure 1: The methylated cytosines are marked by yellow in DNA sequence.

Model Architecture

We select the CNN as our framework, because the property of convolutional layer, that modeling invariant features, allows CNN learns correct "motifs" from DNA sequences. In addition, we add "inception module" into our model so that extract more representative features through kernels of different scales.

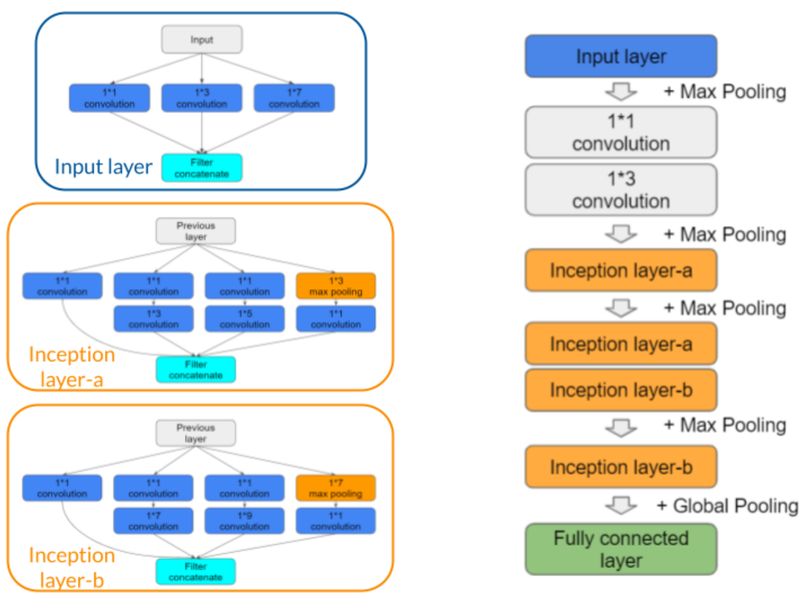

Figure 2: The complete architecture (right) and the detailed modules (left) are shown.

Imbalanced Learning

Our generated dataset is imbalanced in class distribution of histone modifications. In order to alleviate this problem, we introduce label-based stratified mini-batch learning algorithm, proposed by Peng et al. And, we observe apparent progress on performance [2].

Results

We use two metrics to evaluate our results. There are respectively area under the receiver operating characteristics curve (ROC) and area under the precision-recall curve (PRC). To observe how much different input features influence the outcomes. We separately extract features from only DNA sequence, only DNA methylation signal and both by our model, as well as represent them by subscripts in the following tables.

Model	$Inception_{Meth}$	$Inception_{DNA}$	$Inception_{DNA+Meth}$
H3K4me3	0.9214	0.9657	**0.984**
H3K27ac	0.8422	0.8659	**0.9288**
H3K4me1	0.8772	0.8511	**0.9441**
H3K36me3	0.9576	0.8998	**0.9774**
H3K9me3	0.8018	0.9518	**0.9719**
H3K27me3	0.9557	0.909	**0.9915**

Table 1: This table represents performance with AUROC.

Model	$Inception_{Meth}$	$Inception_{DNA}$	$Inception_{DNA+Meth}$
H3K4me3	0.6863	0.83	**0.9006**
H3K27ac	0.6231	0.6662	**0.7862**
H3K4me1	0.7223	0.6694	**0.8565**
H3K36me3	0.8803	0.7859	**0.9415**
H3K9me3	0.2386	0.6865	**0.7987**
H3K27me3	0.8995	0.8497	**0.9834**

Table 2: This table represents performance with AUPRC.

Discussion

In order to observe which input feature provide the most information of crossing cell lines, we visualize the individual feature vectors generated by models, which models extract feature from different input data.

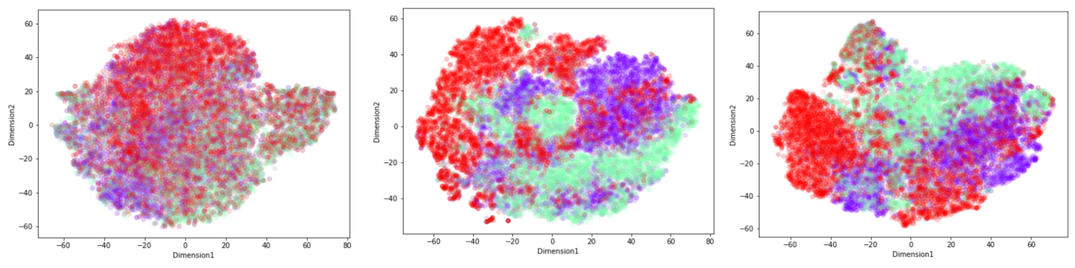

Figure 3: Extract from DNA sequences (left), extract from DNA methylation (middle) and extract from both (right) are shown.

As long as adding input feature of DNA methylation, we clearly observe the model automatically discriminate different biosamples within whole feature vectors. This phenomenon is not observed with the model only using input features of DNA sequences.

Conclusions

In this thesis, we design a novel CNN framework to predict six core histone modifications from DNA sequences and methylation, for improving data imputation in cell line-specific cases. In our model, we introduce an inception module and stratified mini-batch to enhance prediction, and get good performance. According to experimental results, we can observe that model integrating features from DNA sequences and methylation outperforms only using features of DNA sequences. Furthermore, we observe that DNA methylation definitely provides cell line-specific insight for model, by visualization. Thus, we verify that the DNA methylation strengthens the ability of data imputation for cell line-specific cases, with deep-learning model.

Forthcoming Research

- There are many interactions between histone modification and DNA methylation. In the future, we can discuss more about this interesting biological interactions by powerful model.
- Owing to histone is main element of nucleosome, we could investigate the association with nucleosome by deep-learning model in future studies.

References

[1] Howard Cedar and Yehudit Bergman. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304, 2009.

[2] Dunlu Peng, Tianfei Gu, Xue Hu, and Cong Liu. Addressing the multi-label imbalance for neural networks: An approach based on stratified mini-batches. *Neurocomputing*, 435:91–102, 2021.