Prediction of Transcription Factor Binding Sites Using Deep Learning

Rudra Malali[1] Naman Jangid[2] Pranjali Deshmukh[3]
[1,2,3]BE Student
[1,2,3]Department of Computer Engineering
[1,2,3]Zeal College of Engineering and Research, Pune, India

Abstract— The transcription binding sites form the base for knowing the gene regulation and DNA annotation. The prediction of Transcription factor binding sites (TFBSs) is a vital step in genomic study and is carried out by using varied methods like Chip-Seq, Positive-Weight Matrix, Deep Poly and many more but according to studies, the Neural Networks have shown more accurate result. Thus, in our paper we are using a deep learning approach, taking DNA sequence, Chromatin Accessibility, Histone Modification and TATA box sites as inputs to our CNN model, to obtain more accurate TFBSs. During model training, we also used the complementary DNA sequence of the DNA strands in the dataset to achieve better results.

Keywords: Transcription Factor Binding Sites (TFBSs), Convolutional Neural Network, Chromatin accessibility, Histone modification, TATA Box

I. INTRODUCTION

The Transcription is a critical step which regulates gene expression and maintains the cell's identity. Transcription helps knowing which gene will show up its characteristics and regulate the gene expression[1]. The facilitation and repression of transcription depends on the Transcription Factor (TF) that binds at particular cis-regulatory elements in promoter regions of DNA. These regions are called as Transcription Factor Binding Sites (TFBSs), which regulate many genomic characteristics of the DNA [2].

Knowing the importance of TFBSs in genomics many physical methods are implemented since the Human Genome Project, in studying the eukaryotic cells. These methods (like ChIP-seq) are expensive as well as tedious, as they consider only one single cell type and demand tissues [3]. The need of computation methods was thus felt leading many algorithms like Position Weight Matrices (PWMs), Support Vector Machine(SVM), ensemble random forest models and many more to be used. Out of all these, Neural Network shave outperformed all these methods. [4] And more of Convolutional neural network(CNN)[5].

Thus, we have proposed a method which uses the computational advantages of CNN and predict the TFBSs as discussed further in the paper, which is structured as Section II Literature Review, Section III Technical Background, Section IV Result and Analysis and Section V includes Conclusion and discusses the future scope for research.

II. LITERATURE REVIEW

The transcription has a huge importance in understanding the genomics and DNA annotation. Thus a detailed study of binding sites and their context like orientation, motif locations, nucleosome occupancy is done [6]. There is a wide range of research that has been done over different influencing factors for transcription and its application in understanding the immune responses and their characteristics towards target motifs [7], and studying cancer somatic mutation [8]. Having a wide range of application and complexities in experiments the need to computational algorithms was felt.

The methods involved in predicting the TFBSs started with simple methods of using Markov modeling [9]. Then use of Positive Weight matrices was seemed to have decent success [10]. The problems faced in position specific energy matrices (PSEMs) were tried to overcome by using ensemble random forest model [11]. The computation footprinting methods considering ATAC-seq are also implemented in finding TFBSs [12].

With time the use of machine learning approach was taken up by Yuanqizeng and Wuzhong dong in [13] using semi- supervised algorithms. They used a CNN with classification layers and multiple hidden layers to find TFBSs.

The [14] show a use of Convolutional neural network and kernel method in predicting TFBSs and protein homology. Daniel Quangand Xiaohui Xie used FactorNet model to predict binding sites using forward and reverse complement DNA sequence for more accuracy [15].

All these methods used so far shows that considering multiple influencing inputs and using DNA strands complementary reverse sequence in Convolutional Neural Network will help achieve the Transcription binding sites precisely which we have implemented in our model.

III. TECHNICAL BACKGROUND

A. *Flow of Implementation*

The proposed method can be divided as preparing inputs and building model. The dataset of DNA sequences are obtained from [5]. The dataset is in FASTA format with each DNA sequence of length 100bps (base pairs).The probability of finding the DNA motifs in length 100bps of promoters region of DNA is more.

The DNA sequence is used to predict the Chromatin accessibility, using which we generate the Histone modification data. We also use DNA sequence in finding TATA-box motifs using Multiple Bloom Filters. Then all these inputs DNA sequence, Chromatin accessibility, Histone modification data and TATA-box motifs are given to the CNN model to find the TFBSs.

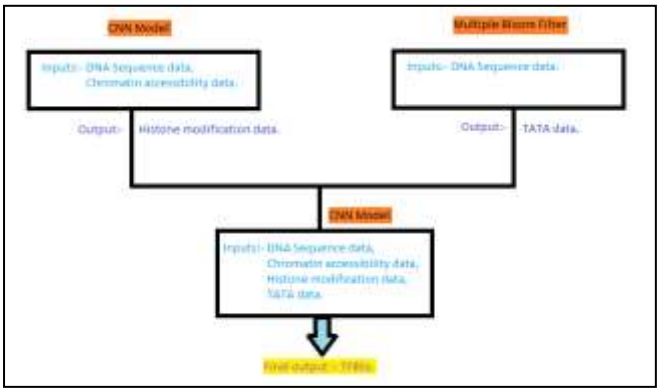

Fig. 1: Flow of system implementation

B. Preparative Stage

The stage one of proposed method starts with predicting the Chromatin Accessibility in DNA sequence using the Deep openness prediction network (Deopen), a hybrid computational model built using deep CNN and three layer FFN (Feed Forward Network)[16].

The Chromatin accessibility data is then further used in predicting Histone Modification using sequence data for varied histone markers. This is achieved through DeepHistone, a DNN (deep neural network) proposed in paper [17], it has used the classification model to extract the Histone modification motifs for varied Histone Markers using DNA module and DNAse module.

The Bloom filters is a pattern matching algorithm that helps us extracting the TATA-box motifs in DNA sequence[18].

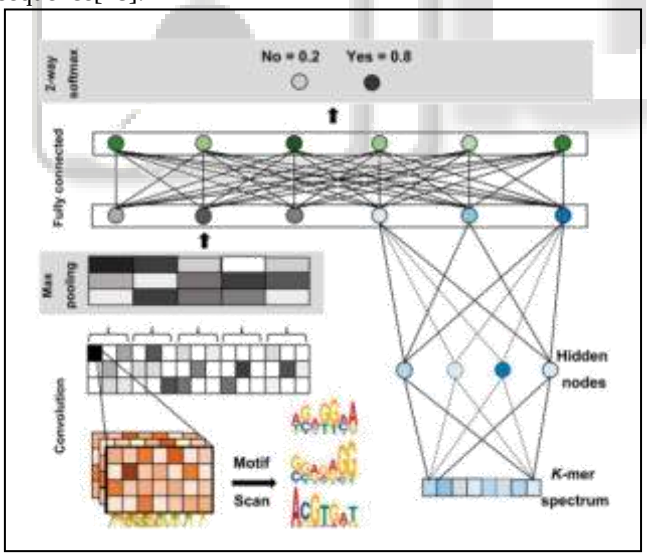

Fig. 2: Deopen Schematic Diagram [16]

C. Convolutional Neural Network (main module)

The Process of prediction TFBSs begins once all the outputs of data input modules are obtained in the required format. The main prediction module is a Convolutional Neural Network, having inputs:-

- DNA sequence and its reverse complementary sequence.
- Chromatin Accessibility
- Histone Modification for particular Histone Marker
- TATA-box motifs.

The main module uses the one hot encoding format for accepting the inputs. It uses forward and reverse DNA sequence to help verify the results coming from both inputs and helps train the model more specifically. The Convolutional network extracts the features and predicts the TFBSs with the depths.

IV. RESULTS AND ANALYSIS

We performed the computation over two different datasets having data variation with respect to sequences and cell types. The accuracy achieved was observed to improve than the previous studies, which are 0.9536 and 0.9707 for the two datasets used.

The results obtained are shown in fig 3, showing the probability of TFBSs over the DNA sequences of 100bps for the given inputs.

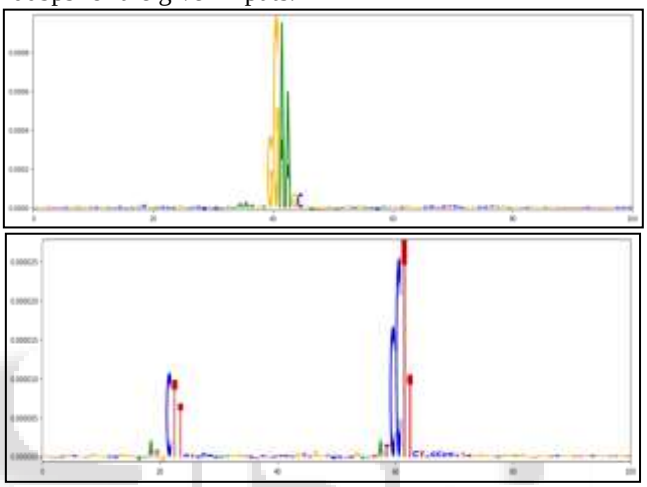

Fig. 3: Results for proposed model

The improvement in results is observed because of considering multiple influencing factors as input and by using complementary DNA sequence for training purpose. The accuracy achieved in by [5] lies around 0.901-0.93AUC.

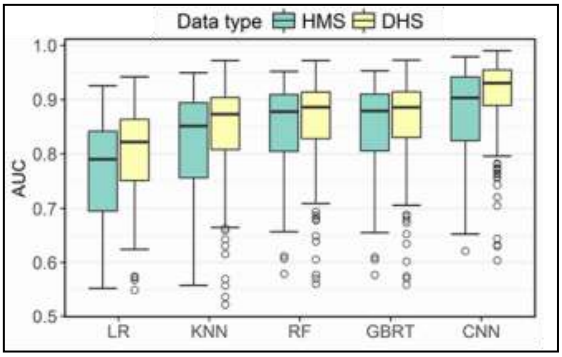

Fig. 4: The comparative analysis of CNN in [5]

The results of paper [19] have precision of 87% and recall of 77% using deconvNet and having single nucleotide resolution. The results we have observed have thus been good to locate the TFBSs in given DNA sequence.

V. CONCLUSION AND FUTURE SCOPE

The transcription forms the bases of gene regulation and defining the functional schema of a gene. We observed that considering the genomic data inputs that influence the Transcription Factor binding sites if took into integrative

structure for building a CNN model helps us achieve more accuracy. Locating TF in DNA promoter regions is more efficient by using TATA-box motif and using Complementary DNA sequences helps better train the model to achieve more precision. The model takes only DNA sequence as its input while the rest of the inputs are extracted by the algorithm itself making it more easy to use.
In the future, we can attempt improving the prediction by integrating other weighted inputs like AUG start codon, nucleosome positions, factors affecting promoters regions, DNA shape and also integrating Neural network with different logistic algorithms.

REFERENCES

[1] Jiannan GUO "Transcription: the epicenter of gene expression",2015, Journal of Zhejiang University SCIENCE B, pp. 409–411.
[2] Yonatan Bilu* and Naama Barkai"The design of transcription-factor binding sites is affected by combinatorial regulation", 2005, Genome, pp. 6:R103.
[3] Peter J. Park, "ChIP–seq: advantages and challenges of a maturing technology", 2009, Nature Reviews Genetics, pp. 669–680.
[4] Hossein Banki-Koshki, S.AliSeyyedsalehi and Fatemeh Zare-Mirakabad "Transcription factor binding sites identification on human genome using an artificial neural network", 2017, Iranian Conference on Electrical Engineering (ICEE), IEEE , pp. 14-17
[5] Fang Jing, Shao-Wu Zhang, Zhen Cao, and Shihua Zhang, "An integrative framework for combining sequence and epigenomic data to predict transcription factor binding sites using deep learning", 2019IEEE/ACM transactions on computational biology and bioinformatics.
[6] Jakub Orzechowski Westholm, Feifei Xu, Hans Ronne and Jan Komorowski, "Genome-scale study of the importance of binding site context for transcription factor binding and gene regulation", 2008BMC Bioinformatics, pp. 484 –498.
[7] Peter S. Askovich*, Stephen A. Ramsey, Alan H. Diercks, Kathleen A. Kennedy, Theo A. Knijnenburg and Alan Aderem, "Identifying novel transcription factors involved in the inflammatory response by using binding site motif scanning in genomic regions defined by histone acetylation.",2017,PLoS One.
[8] Vorontsov, I. E., Khimulya, G., Lukianova, E. N., Nikolaeva, D. D., Eliseeva, I. A., Kulakovskiy, I. V., &Makeev, V. J , "Negative selection maintains transcription factor binding motifs in human cancer", 2016, BMC Genomics.
[9] Rajasekhar Raman & G. Christian Overton, "Application of Hidden Markov Modeling to the Characterization of Transcription Factor Binding Sites.", 1994, Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences HICSS-94.
[10] Gusmao, E. G., Dieterich, C., Zenke, M., & Costa, I. G., "Detection of Active Transcription Factor Binding Sites with the Combination of DNase Hypersensitivity and Histone Modifications", 2014, Bioinformatics,pp. 3143–3151.
[11] Fatemeh BehjatiArdakani, Florian Schmidt, and Marcel H. Schulz, "Predicting transcription factor binding using ensemble random forest models.",F1000Research 7 (2019).
[12] Zhijian Li1, Marcel H. Schulz3, Thomas Look, Matthias Begemann, Martin Zenke and Ivan G. Costa, "Identification of transcription factor binding sites using ATAC-seq.", 2019, Genome biology 20.1, pp. 45-66.
[13] Yuanqi Zeng, Wuzhong Dong , Qingyuan Chen , Yongqing Zhang and Dongrui Gao, "A Transcription Factor Binding Site Prediction Algorithm Based on Semi-Supervised Learning. ", 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing", 2019, IEEE , pp.183-186.
[14] Dexiong Chen, Laurent Jacob and Julien Mairal, "Biological Sequence Modeling with Convolutional Kernel Networks", 2019, Bioinformatics, pp. 3294-3302.
[15] Daniel Quang and XiaohuiXie, "FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data", 2019, Methods 166, pp. 40-47.
[16] Qiao Liu, Fei Xia, Qijin Yin and Rui Jiang, "Chromatin accessibility prediction via a hybrid deep convolutional neural network", 2018, Bioinformatics, pp. 732-738.
[17] Qijin Yin, Mengmeng Wu, Qiao Liu, HairongLv and Rui Jiang, "DeepHistone: a deep learning approach to predicting histone modifications", 2019, BMC genomics, pp. 193-206.
[18] Maleeha Najam, Raihan Ur Rasool, Hafiz Farooq Ahmad,Usman Ashraf, and Asad Waqar Malik. "Najam, Maleeha, "Pattern Matching for DNA Sequencing Data Using Multiple Bloom Filters", 2019, BioMed research international 2019.
[19] Sirajul Salekin, Jianqiu (Michelle) Zhang and Yufei Huang, "Deep learning model for predicting transcription factor binding location at single nucleotide resolution", 2017, IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) , pp.57-60

