Deep Convolutional Neural Network for prediction of DNA accessibility regions of T-cells and epithelial cells for DNase-seq data

Sai Ganesh, Vyshnavi, Shanmukh, Abbas, Yudhisha

saedunu@siue.edu

1. Abstract

DNA accessibility, characterized by open chromatin regions, is a critical determinant of cell identity and function. These accessible regions facilitate the interaction of regulatory elements with the transcriptional machinery, ultimately leading to protein production. Here, we employ Convolutional Neural Networks (CNNs) to decipher the accessibility landscape of DNA. Our CNN models aim to predict accessible regions within DNase-seq data for both T-cells and epithelial cells. This study sheds light on the power of deep learning for deciphering chromatin accessibility patterns, offering insights into cell-specific regulatory programs.

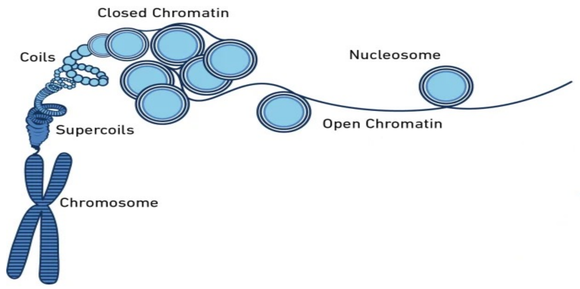
2. Introduction

In recent years, understanding the regulatory mechanisms governing gene expression has been greatly enhanced by advances in the study of DNA accessibility [1]. DNA accessibility, a key determinant of cell identity and function, plays a pivotal role in orchestrating complex regulatory networks within cells . However, deciphering the landscape of accessible chromatin regions has presented a significant challenge to researchers [2].

To address this challenge, we turn to Convolutional Neural Networks (CNNs), inspired by recent advancements in neural network architectures [3]. CNNs offer a powerful tool for unraveling the accessibility landscape of DNA, particularly within DNase-seq data for diverse cell types such as T-cells and epithelial cells . By leveraging insights from these advanced computational models, we aim to elucidate the intricate regulatory mechanisms underlying DNA accessibility .

Ultimately, our research endeavors to deepen our understanding of chromatin biology and its implications for cellular function and disease .

DNA Accessibility :



3. Methodology

3.1 Data

Aspect	Example
Data Type	Bed narrow peak file
Source	ENCODE website (https://www.encodeproject.org/)
Parameters	Assay type: DNase-seq , Biosample: Homo sapiens ,Organ: blood ,Cell: T-cells and epithelial cells ,Analysis: GRCh38 , Read length: 101 for T-cells and 151 for epithelial cells
Size	T-cells: 37 files Epithelial cells: 27 files
Types of Cells	T-cells: Lymphocytes involved in immune response. Epithelial Cells: Form linings of organs, serving protective functions.

3.2 Data Preprocessing

Algorithm: ProcessDNaseSeqData
Input: Bed files of T-cells and epithelial cells
Output: Accessible and Inaccessible files containing DNA sequences of the same length
Step 1: Create two lists, one for storing accessible files and another for inaccessible files.
Step 2: Repeat the following steps for each cell type (T-cells and epithelial cells):

a. Filter DNase-seq data from the ENCODE website based on specified parameters: - Assay type (DNase-seq) - Bio sample (Homo sapiens) - Organ (blood, brain, bodily fluid) - Cell type (T-cells or epithelial cells) - Analysis (GRCh38) - Read length (101 or 151)
b. Download the bed narrow Peak files.
c. Convert each bed narrow Peak file into a nucleotide (ATGCs) file using bed tools.
d. Extract nucleotides from each file and store them separately.
e. Determine an optimal length, X, for cropping sequences to ensure uniform length.
f. Discard sequences shorter than X and trim sequences longer than X.
g. Store the cropped sequences in the accessible file.
h. Create a negative file by subtracting the distance between consecutive accessible regions from the narrow Peak file.
i. Apply the same cropping process to the inaccessible regions.
j. Store the cropped sequences in inaccessible files.

Step 3: Return the accessible and inaccessible files containing DNA sequences of the same length for T-cells and epithelial cells.
Algorithm End

3.3 Hyper Parameter Tuning

Hyperparameters	**Epoch**: Epithelial cell-10, Tcell:10,**Learning Rate**: 0.001,**Optimizer**: Adam,**Loss Function:** Cross-entropy loss, **Dropout Rate:**0.5

4. Network

CV Layer	Network-1
CV1	In_CH=4, Out_CH=96 Kernel_SZ=11, Stride=4
MaxPool1	In_CH=96, Out_CH=96 Kernel_SZ=3, Stride=2
CV2	In_CH=96, Out_CH=256 Kernel_SZ=5, Padding=2
MaxPool2	In_CH=256, Out_CH=256 Kernel_SZ=3, Stride=2
CV3	In_CH=256, Out_CH=384 Kernel_SZ=3, Padding=1
CV4	In_CH=384, Out_CH=384 Kernel_SZ=3, Padding=1
CV5	In_CH=384, Out_CH=256 Kernel_SZ=3, Padding=1
MaxPool5	In_CH=256, Out_CH=256 Kernel_SZ=3, Stride=2
Fc1	In_CH=256*4/3, Out_CH=2048
Fc2	In_CH=2048, Out_CH=2048
Fc3	In_CH=2048, Out_CH=2

CV Layer	Network-2
CV1	In_CH=4, Out_CH=96 Kernel_SZ=11, Stride=4, pad=0
CV2	In_CH=96, Out_CH=96 Kernel_SZ=1
CV3	In_CH=96, Out_CH=96 Kernel_SZ=1
MaxPool3	Kernel_SZ=2, Stride=2
CV4	In_CH=96, Out_CH=256 Kernel_SZ=11, Stride=4, pad=2
CV5	In_CH=256, Out_CH=256 Kernel_SZ=1
CV6	In_CH=256, Out_CH=256 Kernel_SZ=1
MaxPool6	Kernel_SZ=3, Stride=2
CV7	In_CH=256, Out_CH=384 Kernel_SZ=3, Stride=2, pad=1
CV8	In_CH=384, Out_CH=384 Kernel_SZ=1
CV9	In_CH=384, Out_CH=384 Kernel_SZ=1
Fc1	In_CH=384*1, Out_CH=2048
Fc2	In_CH=2048, Out_CH=2048
Fc3	In_CH=2048, Out_CH=2

5. Results

Cell Type	T-cell	Epithelial cell
Network-1	85%	86%
Network-2	81%	80%

6. Conclusion/Limitations

- The study effectively achieved a reasonable level of accuracy in predicting DNA accessibility regions utilizing DNase-seq data for T-cells and epithelial cells.
- The investigation revealed that increasing the depth of the model did not consistently enhance predictive performance, underscoring the importance of striking a balance between model complexity and depth.
- The model's applicability is confined to DNase-seq data originating from T-cells and epithelial cells.
- We have undertaken the necessary step of trimming the original DNA sequence to conform to the input requirements of the CNN model, as is standard practice across existing CNN models. However, this process inevitably results in the loss of certain portions of the original data.
- The lack of explainable AI attributes within the model restricts the interpretability of predictions, potentially limiting insights garnered from its application.

7. References

[1] Bogan, S. N., Strader, M. E., & Hofmann, G. E. (2023). Associations between DNA methylation and gene regulation depend on chromatin accessibility during transgenerational plasticity. BMC biology, 21(1), 149.

[2] Hoffman, G. E., Bendl, J., Girdhar, K., Schadt, E. E., & Roussos, P. (2019). Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. Nucleic acids research, 47(20), 10597-10611.

[3] Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research, 26(7), 990-999.