The Identification of Transcriptionally Active Regions

Dante Miller
Mill8622@morris.umn.edu
Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA

Abstract

Current methods for identifying transcriptionally active regions are expensive, arduous, and require large portions of cells. Another issue is their inefficiency in distinguishing individual transcriptionally active regions from large sites of transcriptionally active regions. Recent research by Tripodi et al. [7] utilized ATAC-seq and recurrent neural networks in classifying open chromatin regions as to whether the regions are transcriptionally active or not. Through the identification of transcriptionally active regions, mechanisms for diseases such as cancer can be better understood. This paper provides an overview of biological concepts and an implementation with background information on the sub-techniques in the overall algorithm.

Keywords: Transcription, open chromatin regions, computational biology, genome-wide chromatin accessibility, nascent transcription, machine learning

1 Introduction

Transcription and translation are two of the main processes for the flow of genetic information relating to the health and physical well-being of an organism. Mutations are changes in a DNA sequence either due to mistakes during the transcription process, translation process or caused by environmental factors which can also occur in cohort. Diseases are caused by mutations in a genes, influenced by environment factors or damaged chromosomes which can also occur in cohort. *Open chromatin regions* (OCRs) are regions in the DNA where transcription can potentially take place. Recent studies have shown that a series of OCRs are associated with mechanisms in human diseases such as cancer [8]. The identification of mechanisms behind diseases is crucial in understanding what mechanisms are associated with specific diseases. Potential treatments and medications for specific diseases can be determined by understanding disease mechanisms. The identification of mechanisms for specific diseases can be better understood through studies about chromatin accessibility and *transcriptionally active regions* (TARs) [6].

TARs are regions where the transcription process is taking place. GRO-seq and PRO-seq are nascent transcription methods that are currently used in identifying TARs or nascent transcripts which are mentioned in Section 2.1. These methods are expensive, arduous, and require large portions of cells [7]. Another issue is their inefficiency in distinguishing individual TARs from large sites of TARs.

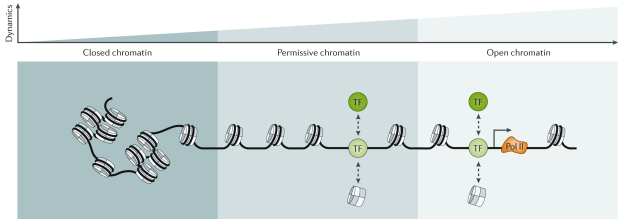

Figure 1. A visual image on the different types of chromatin accessibility. The three blue areas of the image left from right represent closed, permissive, and open accessibility. [5]

Only a portion of OCRs are transcriptionally active which is why it is important to classify OCRs as to whether the regions are transcriptionally active or not [7]. Tripodi et al. implemented a method that is simpler than other methods for identifying TARs and can accurately classify OCRs as to whether the regions are transcriptionally active or not while accomodating a wide range of cell types and counts. This implementation consists of the use of ATAC-seq and *recurrent neural networks* (RNNs). ATAC-seq is a method that identifies OCRs and accommodates all the current problems present in nascent transcription methods, but it can not identify TARs. The application of RNNs to the results of ATAC-seq is a novel way of classifying OCRs as to whether the regions are transcriptionally active or not due to the sequential nature of the data produced by ATAC-seq and the OCRs potentially inducing some form of signal [7].

In this paper, section 2 provides an overview of the transcription process. Background information is provided in relation to RNNs, DNA sequencing, ATAC-seq, and nascent transcription methods (GRO-seq/PRO-seq) to better understand how the data was obtained in the study by Tripodi et al. Section 3 provides a brief overview of the algorithm implementation and the results of the study by Tripodi et al. Section 4 wraps up the conclusions made from the study by Tripodi et al. and future research that can be done in identifying TARs.

2 Background

This section provides background information relating to transcription to better understand the study by Tripodi et al. Background information is also provided in regards to

RNNs, ATAC-seq, and nascent transcription methods (GRO-seq/PRO-seq) to better understand the methods used in the study and how the data were obtained. An overview of the data or short read runs is also provided.

2.1 Transcription

The transcription process involves chromosomes, a long molecule consisting of the genetic information within an organism. Proteins are a form of molecules involved in the structure and functions of cells. Proteins are formed out of smaller units known as amino acids. In the chromosome, there are special regions known as genes that contain instructions for which specific amino acid to use in constructing the protein or producing molecules to help the cells assemble the protein. DNA, the building block of genes, is composed of organic molecules such as adenine (A), cytosine (C), guanine (G), and thymine (T).

The transcription process is managed by an enzyme called called *RNA polymerase* (RNAP). The purpose of the RNAP enzyme is to copy the DNA at the specific gene into a molecule called a *messenger RNA* (mRNA). The mRNA molecule is used in the translation process where the genes are then translated into proteins. Promoters are regions that define the starting point of where the RNAP enzyme begins the transcription process. The RNAP enzyme pulls apart the gene it is transcribing into two strands called the template and non-template strands. The RNAP enzyme then walks along the template strand while synthesizing the mRNA as it zips the template and non-template strands back together. The mRNA before the transcription process terminates is known as a *nascent transcript*, an earlier version of the mature mRNA. Once the RNAP enzyme reaches the end of the gene, the transcription process terminates.

Transcription factors are proteins involved in the transcription or translation processes. Chromatin is the material that makes up chromosomes. In Figure 1, there are several physical forms of chromatin that dictate how accessible the DNA at those regions are to transcription factors and the RNAP enzyme. OCRs allow transcription factors and the RNAP enzyme to access the DNA to transcribe the genes into an mRNA. Closed chromatin regions do not allow transcription factors and the RNAP enzyme to access the DNA to transcribe the genes into an mRNA. Permissive chromatin regions are permissive of transcription factors.

2.2 Recurrent Neural Networks

It is beneficial to understand feedforward neural network (FNNs) before grasping the concept of RNNs [3].

2.2.1 Structure. Machine learning is the study of statistical models that automatically improve themselves through training. Artificial neural networks (ANNs) are a subset of machine learning consisting of computing systems inspired by the physical processes observed in human brains [3].

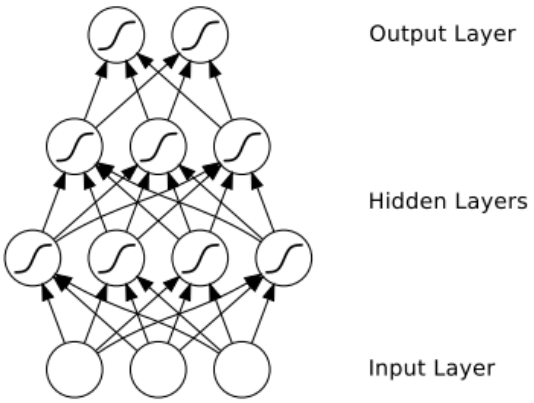

Figure 2. Feedforward Neural Network. [3]

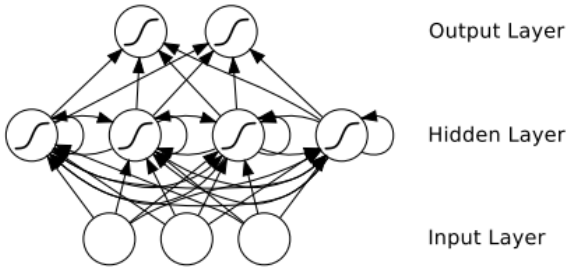

Figure 3. Recurrent Neural Network. [3]

FNNs are one of the many forms of ANNs. In Figure 2, the FNN is organized into an input layer, hidden layers, and an output layer. A node is a computational unit that receives values from the previous nodes, performs computation and then outputs the computation value to the nodes in the next layer or as the final output value. In each layer, there is a certain number of nodes. The transmission of values between nodes in different layers is done through edges. Each of the edges contains a weight that decides how much influence the input value will have on the output value.

2.2.2 Activation Function. Each node in the hidden and output layers consists of an activation function to determine the output of the node. It should be mentioned that each node is part of a calculation in determining the expected output. Activation functions are typically non-linear as it allows FNNs to map the inputs that vary non-linearly to the expected outputs. In Figure 2, the activation function is represented by the lines in the nodes. Equation (1) represents the computation for each node. The activation function (f) is applied to the sum of all previous node output values received (v_i) by the current node multiplied by edge weights (w_i).

$$f(\sum_{i=1}^{t} v_i * w_i) \quad (1)$$

2.2.3 Error Rate. The error rate is the distance between the expected outputs of the dataset provided to the FNN and the results produced by the FNN.

2.2.4 Forward Propagation and Backpropagation. Forward propagation and backpropagation are two processes utilized by FNNs. Forward propagation is a process in FNNs that consists of mapping input values to an output value. Backpropagation is the process in an FNNs that consists of tuning the weights to lower the error rate.

2.2.5 Training. To better understand forward propagation and backpropagation, this section provides a detailed overview of the training process.

A given dataset is randomly split into three different subsets, such as the training, validating, and testing. The training dataset is provided to the FNN. The FNN starts by assigning random weights and then utilizes the forward propagation process to determine the first set of predicted outputs. An error rate is then determined based on the set of predicted outputs and expected outputs. The weights in the hidden and output layers are then adjusted through the backpropagation process. Epochs is the number of times the training dataset is ran through the FNN. Batches are the number of training samples ran per epoch. Iterations are the number of batches needed per epoch. The forward propagation and backpropagation processes are run until the error rate reaches a sufficient value or the number of iterations have been completed. If a validation dataset is provided, then an unbiased evaluation is performed to tune the hyperparameters of the FNN. The training process is then repeated with the new set of hyperparameters.

After the training process, the testing dataset is used to evaluate the performance of the results produced by the FNN.

2.2.6 Differences between FNNs and RNNs. RNNs were derived from FNNs but instead of handling individual data points, RNNs handle sequential data. RNNs allow previous data observations to affect the next output. In Figure 3, notice that in the hidden layers, there are loops. The loops represent how the inputs from each previous observation have an effect on the input in the next observation.

2.2.7 Vanishing Gradient Problem. RNNs suffer from an issue known as the vanishing gradient problem. The vanishing gradient problem is an issue where the weights of earlier layers can not be tuned during backpropagation. Gated recurrent unit (GRU) is a mechanism developed for solving the vanishing gradient problem through the use of an update gate, reset gate, and a hidden state. A hidden state stores information from previously processed sequenced items. An update and reset gate dictate the information that is allowed to the output of a node while determining the relevance of information to the output of a node. A bidirectional RNN is the use of two RNNs, one taking the input in a forward direction, and another taking the input in a backwards direction. A bidirectional GRU is when each of the RNNs from the bidirectional RNN implement a GRU.

2.3 DNA Sequencing

DNA sequencing is a technique for determining reads or the exact sequence of the organic molecules in a DNA molecule. Multiple copies of DNA fragments are made through a process known as amplification. Short read runs are a set of raw sequences after amplifying and then sequencing the copied short fragments of the DNA.

2.4 ATAC-Seq

ATAC-seq is a method that determines the accessibility of chromatin, making it a useful method in providing information relating to the genetic mechanisms of diseases [2]. This section provides an overview of the ATAC-seq method and of a pipeline for processing and interpreting the short reads runs.

2.4.1 Preprocessing. Before covering the preprocessing phase of the ATAC-seq short read runs, it is important to understand how ATAC-seq short read runs were obtained. Sequencing adapters are short single strands of synthetic DNA or RNA. The ATAC-seq method utilizes an enzyme called Tn5 transpose to insert sequencing adapters at accessible regions of the genome. The set of short DNA fragments are then amplified and sequenced. The short read runs are then stored in a database with annotations about where the short read runs are from, which study the DNA was sequence for and what species the short read runs are from.

Furthermore, the preprocessing phase utilizes the short read runs produced by ATAC-seq. The first step consists of downloading the short read runs and attached annotations. Afterwards, the quality of the short read runs and the presence of sequencing adapters are determined. The short read runs are then trimmed through the removal of the low-quality short read runs and sequencing adapters.

2.4.2 Mapping and Filtering Mapped Reads. A reference genome is a database on one idealized individual organism of a species. In the mapping phase, the trimmed reads produced by the preprocessing phase are mapped to the reference genome of interest through a mapper. A mapper takes the reads and reference genome of interest and then matches the reads to the most similar regions on the reference genome of interest. After mapping the reads to the reference genome of interest, filters are imposed to remove uninformative reads, duplicate reads, low mapping quality, and reads that are not properly paired. The insert size, the length of the space between the sequencing adapters is also

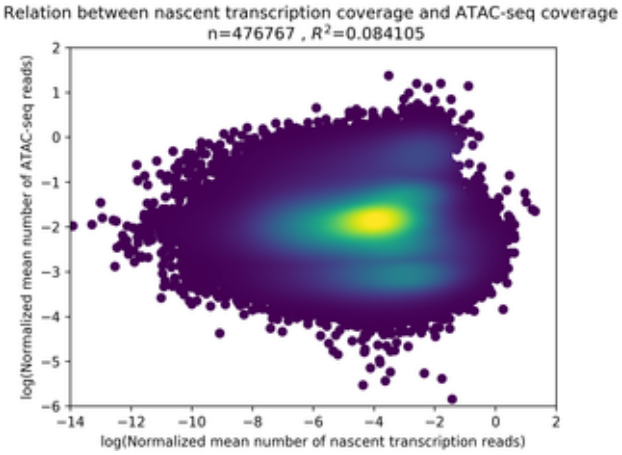

Figure 4. Plot of the relationship between accessibility and nascent transcription. The number of mapped reads for ATAC-seq and nascent transcription are transformed through logarithmic scale. [7]

checked to determine the quality of the reads produced by ATAC-seq.

2.4.3 Peak Calling. The peak calling phase is important in identifying potential OCRs. Model-based Analysis of ChIP-Seq (MACS) is a peak calling method for identifying peaks of the distribution of mapped reads. Each of the regions on the reference genome contains a set number of mapped reads. The distribution of mapped reads on the reference genome of interest can be determined for both strands of DNA. A distribution can be determined based on the two distributions obtained earlier. Regions that are statistically significant are denoted as peaks or OCRs.

2.5 Nascent Transcription

The nascent transcription methods (GRO-seq/PRO-seq) are current tools used for mapping the location and quantity of actively transcribing RNA polymerase at specific sites in the genome [1]. This section provides an overview of nascent transcription methods (GRO-seq/PRO-seq) and their pipelines.

2.5.1 Preprocessing. Before covering the preprocessing phase, we examine how the sequenced reads were obtained. The nascent transcription methods (GRO-seq/PRO-seq) first halt transcription in order to access the nascent transcript molecules before they becomes mature MRNA molecules. The nucleus, a key structure of cells with the purpose of controlling and regulating the activities of cells, is isolated from a cell population of interest. Organic molecules labeled with tags are added. Transcription is then restarted resulting in labeled nascent transcript molecules. The nascent transcripts are isolated, amplified, and sequenced to reveal the location and quantity of RNA production at specific sites in the genome.

Furthermore, the pipeline for the nascent transcription methods is similar to the ATAC-seq method. The sequenced reads are trimmed through the removal of low-quality reads and sequencing adapters. The trimmed reads are then mapped to the reference genome of interest. Filters similar to the ATAC-seq filters are then imposed to check the quality of the mapped reads.

2.5.2 Identifying Transcriptionally Active Open Chromatin Regions. FStitch and Tfit are two current tools used for identifying TARs within a nascent transcription experiment. Hidden Markov model is a model for the purpose of predicting a sequence of hidden variables from observed variables. Logistic regression model is a model for the purpose of predicting a binary outcome based on the observations of a dataset. Finite mixture model is a model used to classify observations into classes. FStitch uses hidden Markov model and logistic regression model to classify regions with large numbers of mapped reads as to whether they are transcriptionally active or not. Tfit used finite mixture models in identifying TARs. FStitch can not distinguish individual nascent transcript in densely transcribed regions and Tfit identifies individual nascent transcripts. The results produced by the FStitch and Tfit are used to train, validate, and test the RNNs in the implementation by Tripodi et al. For more details on Tfit and FStitch see [4].

3 Implementation

This section provides an overview of the implementation of the method provided by Tripodi et al. and the results produced from their study.

3.1 Data Collections

In the data collections section, Tripodi et al. used ATAC-seq and nascent transcription (PRO-seq/GRO-seq) short-read runs provided by different labs through an international public repository known as Gene Expression Omnibus (GEO), a database that consists of data on gene expression and DNA sequencing. Cell lines are cell population types that keep proliferating instead of having a limit on the number of times they can multiply. The different labs used cell lines that represented diseases such as lung cancer, colon cancer, leukemia, invastive ductal cancer, and prostate cancer, which are easier to obtain than taking primary cells from patients with such diseases. There were nine sets of short-read runs selected under the criteria of the ATAC-seq and nascent transcription (PRO-seq/GRO-seq) having matching cell types and diseases.

3.2 Data Preprocessing

The data preprocessing stage in the implementation is similar to what is described in Sections 2.4.1 and 2.5.1. The human

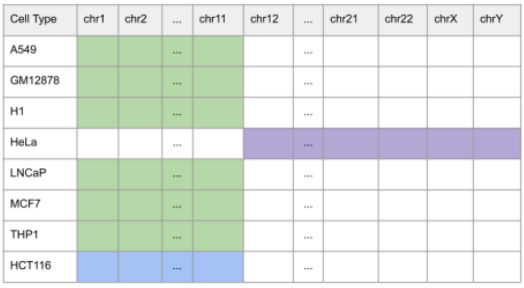

Figure 5. Training, validation, and testing sets. [7]

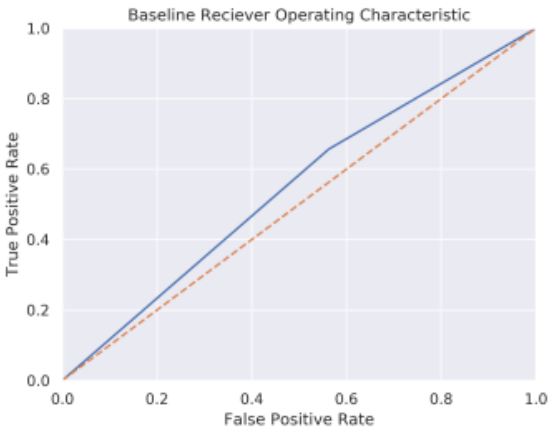

Figure 6. Baseline method ROC plot. [7]

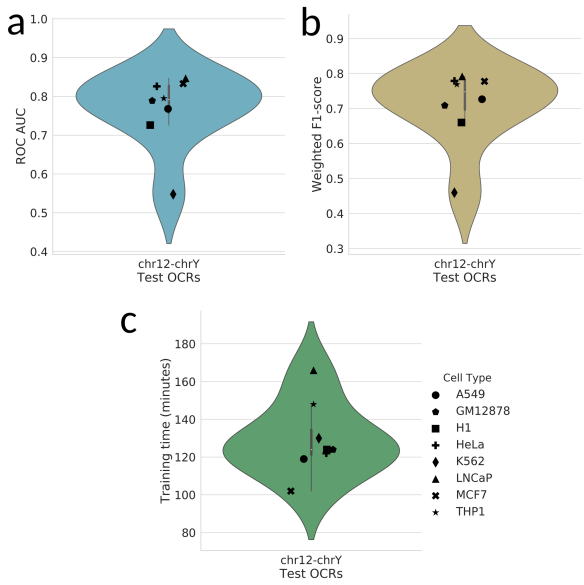

Figure 7. Classifier performance of RNNs. The blue diagram, brown diagram, and green diagram represent the ROC AUC score, F1-score, and training length for each of the RNN models. [7]

reference genome used in Tripodi et al.'s study is GRCh38. Approximately around a half a million OCRs were identified from ATAC-seq pipeline. FStitch and Tfit were performed on the nascent transcription (PRO-seq/GRO-seq) short read runs. Approximately 29 percent of the OCRs were labeled as transcriptionally active. In Figure 4, it can be noticed that there is a low positive correlation between the mean number of reads for ATAC-seq and nascent transcription (PRO-seq/GRO-seq) short read runs for each OCR. This implies that mean number of ATAC-seq reads can not be the sole feature in classifying OCRs as to whether the regions are transcriptionally active or not. The sequence for each OCR may be a useful feature in classifying OCRs as to whether they are transcription active or not. Sequence features may be critical to the transcription process as transcription factors bind to specific regions of a DNA sequence and regions such as promoters occur in a specific position in a DNA sequence [7].

3.3 Data Encoding

An encoding is a technique for converting categorical variables from data into numerical values. Tripodi et al. developed an encoding to condense each OCR into a vector encoding with the minimum loss of information about it. The encoding consisted of sequence and signal. The sequence is mapped to a vector encoding and then concatenated with the signal of each OCR, which is the number of mapped ATAC-seq reads divided by the total number of mapped reads in the millions.

3.4 Classifiers

A RNN model was developed to classify the OCRs represented as ATAC-seq peaks in the signal and sequencing encoding. Keras, an open source Python library for developing and evaluating deep learning models, was used to developed the RNN model. The RNN implemented a bidirectional GRU as the ATAC-seq signal and sequence can be read in either direction.

3.5 Results

A baseline is the result of a very basic model for solving a problem. A kernel density estimator is a method that estimates the probability distribution of the data points it is applied to. A kernel defines the shape of the function used to smooth the data points. A Gaussian kernel is a kernel with the shape of a normal distribution curve. The baseline method used was a kernel density estimator with a Gaussian kernel determine the distribution of mapped ATAC-seq reads per OCR labeled as transcriptionally active or not. Odds ratio is a method that quantifies the association between two

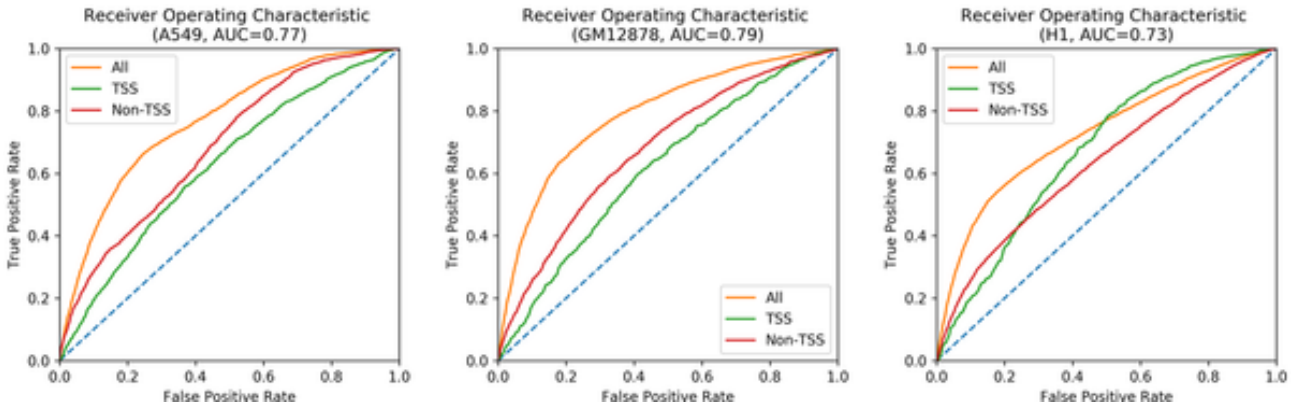

Figure 8. Error analysis using ROC curve for comparing predicted vs actuals. The orange line, green line, and red line represent all OCRs, TSS OCRs, and non-TSS OCRS, respectively. [7]

events. The two events are whether an OCR is transcriptionally active or not. Odds ratio was used to determine whether an OCR overlapped transcription.

1. True positive (TP) is the number of OCRs predicted as transcriptionally active that actually are
2. False negative (FN) is the number of OCRs predicted as not transcriptionally active that actually are
3. True negative (TN) is the number of OCRs predicted as not transcriptionally active that actually are not
4. False positive (FP) is the number of OCRs predicted as transcriptionally active that actually are not

F1-score is the measure of the models prediction accuracy. The range of the F1-score is between 0 and 1. The formula for F1-score can be noticed in Equation (3).

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3}$$

The true positive rate (TPR) and false positive rate (FPR) are used to measure the models ability in predicting the actual categories correctly and incorrectly. The TPR would be the number of OCRs classified correctly as transcriptionally active and actually are. The formula for the TPR can be noticed in Equation (4).

$$TRP = \frac{TP}{TP + FN} \tag{4}$$

The FPR would be the number of OCRs classified as transcriptionally active and actually are not. The formula for the FPR can be noticed in Equation (5).

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

Figure 6 visualizes the trade-off between the TPR and FPR through a receiver operating characteristic curve (ROC curve) which visualizes the performance of a classification model. The dotted line represents random guessing in classifying the OCRs. As the TPR increases, the FPR increases at the same rate. What can be noticed is that the baseline model is similar to the dotted line. The best possible model would be where the TPR is at 1 and the FPR is at 0. The area under the curve (AUC) is a score that can be used to aggregate the performance of the model between the range of 0 and 1. The larger the AUC score is, the better the model is able to classify correctly. The AUC score for the dotted line and the best possible model are 0.5 and 1, respectively. The baseline method generated an F1-score of 0.550 and an ROC AUC score of 0.554 on a 10 percent of the identified randomly selected OCRs.

The RNN was trained and validated eight times. Each cell type was tested in a similar process noticed in Figure 5 except for the HCT116 cell type which was only used as the validation set. In each model, a cell type is not used in any part of the modeling process. The green observations are used to train the model. The purple observations are used to test the RNN.

Figure 7 shows some of the results after training, validating, and testing the RNNs. What can be noticed is that the majority of the cells except for K562 have a ROC AUC higher than 70 percent and an F1-score higher than 65 percent. The training times took over 100 minutes for each of the RNNs. Transcription start site (TSS) is the location where the first DNA nucleotide is transcribed into RNA. TSS only represent a small fraction of all TARs. In Figure 8, it can be noticed that the RNNs performed worse on TSS regions compared to non-TSS regions.

4 Conclusion

TARs can be classified accurately throught the use of ATAC-seq and RNNs. The benefits of ATAC-seq can be maximized by RNNs to better understand disease mechanisms in future studies. Future studies can be done on improving the performance of the RNNs by accounting for the differences in dataset quality, including more information relating to the signal and incorporating annotations or other input data.

Acknowledgments

I would like to show my appreciation to Dr. Elena Machkasova for her guidance and continuous support over the literature review process. I would like to also thank Dr. Peter Dolan, Mitchell Finzel, and Chineng Vang for their feedback over the literature review process.

References

[1] Chris Benner. 2019. Homer Software and Data Download. http://homer.ucsd.edu/homer/ngs/groseq/groseq.html

[2] Lucille Delisle, Maria Dolye, and Florian Heyl. 2021. *ATAC-Seq data analysis (Galaxy Training Materials).* https://training.galaxyproject.org/training-material/topics/epigenetics/tutorials/atac-seq/tutorial.html

[3] Alex Graves. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks.* Studies in Computational Intelligence, Vol. 385. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-24797-2

[4] Margaret Gruca, Michael Gohde, and Robin Dowell. [n.d.]. Annotation Agnostic Approaches to Nascent Transcription Analysis: Fast Read Stitcher and Transcription Fit. http://dna.colorado.edu/assets/pdf/GrucaMethodsPreprint.pdf

[5] Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. 2019. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* 20, 4 (April 2019), 207–220. https://doi.org/10.1038/s41576-018-0089-8

[6] Hannah J. Perrin, Kevin W. Currin, Swarooparani Vadlamudi, Gautam K. Pandey, Kenneth K. Ng, Martin Wabitsch, Markku Laakso, Michael I. Love, and Karen L. Mohlke. 2021. Chromatin accessibility and gene expression during adipocyte differentiation identify context-dependent effects at cardiometabolic GWAS loci. *PLOS Genetics* 17, 10 (Oct. 2021), e1009865. https://doi.org/10.1371/journal.pgen.1009865

[7] Ignacio J. Tripodi, Murad Chowdhury, Margaret Gruca, and Robin D. Dowell. 2020. Combining signal and sequence to detect RNA polymerase initiation in ATAC-seq data. *PLOS ONE* 15, 4 (April 2020), e0232332. https://doi.org/10.1371/journal.pone.0232332

[8] Jiayin Wang, Liubin Chen, Xuanping Zhang, Yao Tong, and Tian Zheng. 2021. OCRDetector: Accurately Detecting Open Chromatin Regions via Plasma Cell-Free DNA Sequencing Data. *International Journal of Molecular Sciences* 22, 11 (May 2021), 5802. https://doi.org/10.3390/ijms22115802