AI Models to Analyze ATAC-Seq Data of Human T-Cell and Lymphoblast Cells for DNA Accessibility

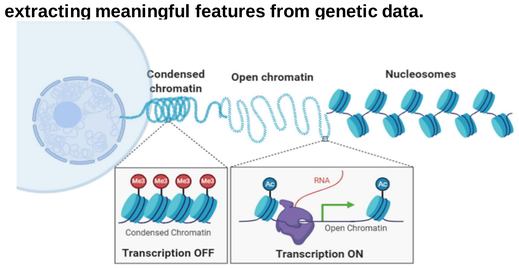
Caleb Bayles | cbayles@siue.edu **Zachery Linscott** | zlinsco@siue.edu

Guided by: **Dr. Manas Jyoti Das**

Southern Illinois University Edwardsville

Abstract

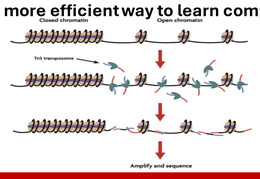
- **DNA accessibility, determined by open and closed regions of the chromatin structure, is an essential detail of cellular biology, and establishes the process of transcriptional regulation.**
- **CNNs have become invaluable in gaining further insight into the regulatory mechanisms of the genome, thanks to their ability to extract meaningful features from large sets of genetic data in a relatively short amount of time.**
- **We utilized CNNs in conjunction with genetic data obtained through the ATAC-seq technique to classify whether sequences exist within an accessible or inaccessible region of the genome.**
- **Our models, trained on T-cells and lymphoblasts, proved to be successful with both high accuracy and precision, providing insight on the capabilities of deep learning in extracting meaningful features from genetic data.**



Source: https://www.frontiersin.org/articles/10.3389/fcell.2021.653077/full

Introduction

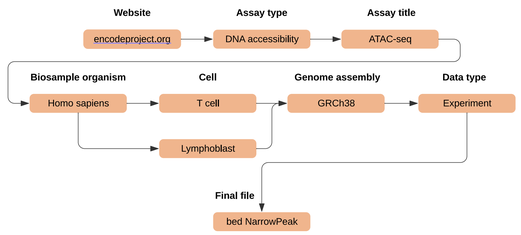
- **We largely have CNNs to thank for modern advances in genetics since they have made it possible to explore complex phenomena such as:**
  - **The impact of mutations on DNA accessibility and transcriptional regulation [1]**
  - **The precise study of chromatin dynamics within a multitude of cancer cells [1]**
- **DNA accessibility is one of the fundamental analyses that can be done given a DNA sequence.**
  - **In this study we study the accessible and inaccessible regions of T-Cell and Lymphoblast cells.**
  - **The cell types are mainly chosen because of the availability of data.**
- **We are mainly considering ATAC-Seq data in our study because it provides information on open chromatin regions [2] better than DNA-Seq.**
- **The AI model plays an important role in analyzing any data. In this study we are employing two different networks:**
  - **AlexNet [3] and Network in Network (NiN) [4].**
  - **AlexNet is a deep architecture, allowing it to capture complex patterns in data.**
  - **NiN is more computationally efficient, and NiN's micro-filters introduced a more efficient way to learn complex features.**



Source: https://www.activemotif.com/blog-atac-seq

1. Dataset

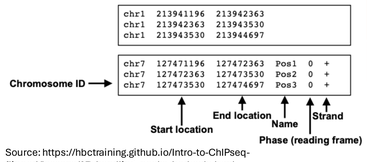
Data Filtering Process



GRCh38 Reference Genome

- We used bedtools, a package for retrieving sequences from the accessible and inaccessible regions of the T-cell and lymphocyte assays.
- Reference genomes serve as a specific representation of an organism's genetics. [5]
- GRCh38 represents the genetics of Homo sapiens. [5]

BED File Format



Source: https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/07_handling_peaks_bedtools.html

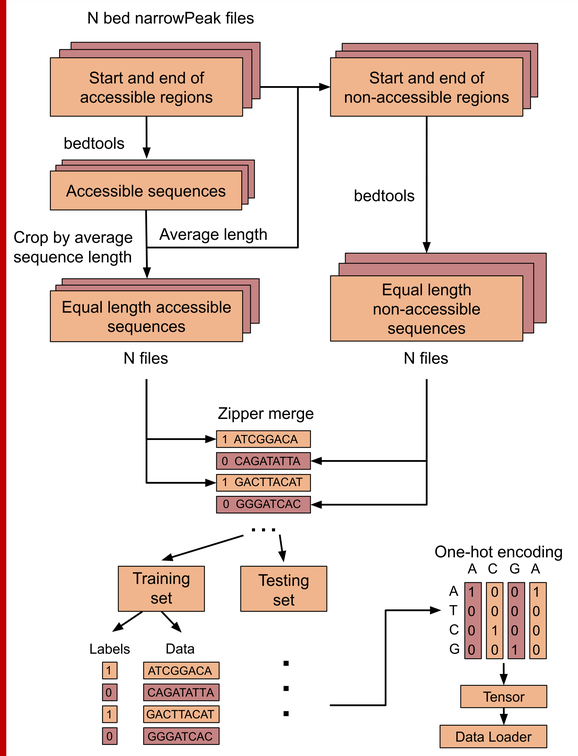
Final Data Size

	# bed narrow peak files	Train Set Size	Test Set Size	Validation Set Size	Mean Sequence Length
T-Cells	42	3038782 sequences	1012928 sequences	595980 sequences	726
Lymphoblasts	56	2626360 sequences	656590 sequences	172788 sequences	678

T-Cells total sequence count: 4,647,690

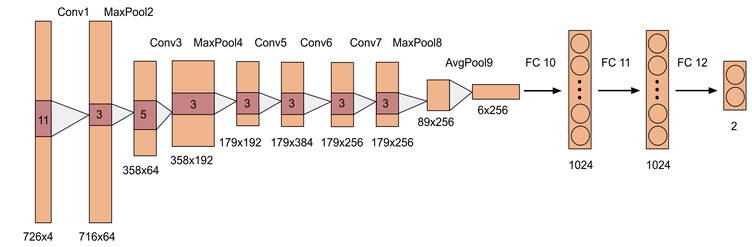
Lymphoblasts total sequence count: 3,455,738

2. Data Preprocessing

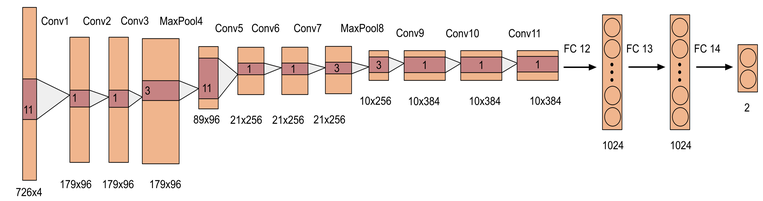


3. Neural Network Architectures

AlexNet

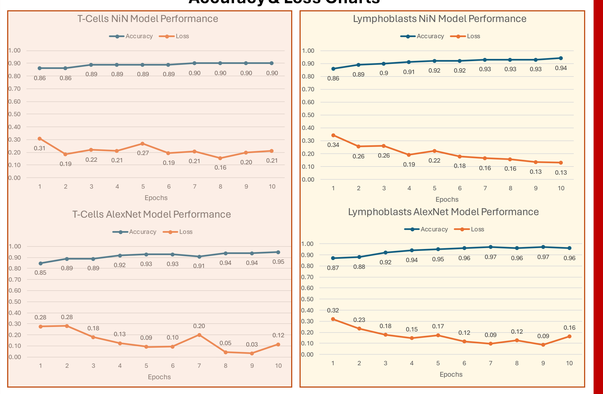


NiN



4. Results

Accuracy & Loss Charts



Confusion Matrices

T-Cells NiN

N = 595980	Predicted 0	Predicted 1	
Actual 0	273152	24838	297990
Actual 1	32995	264995	297990
	306147	289833	

Accuracy: 90%, Precision: 91%

Lymphoblasts NiN

N = 172788	Predicted 0	Predicted 1	
Actual 0	81307	5087	86394
Actual 1	4912	81482	86394
	86219	86569	

Accuracy: 94%, Precision: 94%

T-Cells AlexNet

N= 595980	Predicted 0	Predicted 1	
Actual 0	286708	11282	297990
Actual 1	27277	270713	297990
	313985	281995	

Accuracy: 95%, Precision: 96%

Lymphoblasts AlexNet

N = 172788	Predicted 0	Predicted 1	
Actual 0	84995	1399	86394
Actual 1	5279	81115	86394
	90274	82514	

Accuracy: 96%, Precision: 98%

Conclusion/Limitation

- **Both of our models, trained on T-cells and lymphoblasts, successfully classified regions of DNA from our validation dataset as accessible or non-accessible with high accuracy and precision.**
- **In a different study, we ran tests on T-cells with similar model complexities, sequenced with DNA-seq. We observed the accuracy to be around 83%. This indicates that ATAC-seq may be better for DNA accessibility testing.**
- **Our models have limitations however:**
  - **We were not using explainable AI, but rather a black box model. As a result, it is not possible for us to reconstruct how exactly the models perform accessibility classification.**
  - **It is unknown whether these models would perform well on other datasets. They were trained on a very specific dataset and biological samples and have not been tested for their generalization abilities.**

References

[1] Kamil Wnuk, Jeremi Sudol, Kevin B. Givechian, Patrick Soon-Shiong, Shahrooz Rabizadeh, Christopher Szeto, Charles Vaske "Deep Learning Implicitly Handles Tissue Specific Phenomena to Predict Tumor DNA Accessibility and Immune Activity."
bioRxiv 229385; doi: https://doi.org/10.1101/229385.

[2] Joan Candela-Ferre, Borja Diego-Martin, Jaime Pérez-Alemany, Javier Gallego-Bartolomé, "Mind the gap: Epigenetic regulation of chromatin accessibility in plants," *Plant Physiology*, Volume 194, Issue 4, April 2024, Pages 1998-2016, https://doi.org/10.1093/plphys/kiae024.

[3] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E. (2017-05-24). "ImageNet classification with deep convolutional neural networks". doi:10.1145/3065386

[4] Lin, Min; Chen, Qiang; Yan, Shuicheng. (2013-12-16). "Network in Network." doi: https://doi.org/10.48550/arXiv.1312.4400

[5] Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS, Kremitzki M, Magrini V, Markovic C, McGrath S, Steinberg KM, Auger K, Chow W, Collins J, Harden G, Hubbard T, Pelan S, Simpson JT, Threadgold G, Torrance J, Wood JM, Clarke L, Koren S, Boitano M, Peluso P, Li H, Chin CS, Phillippy AM, Durbin R, Wilson RK, Flicek P, Eichler EE, Church DM. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017 May;27(5):849-864. doi: 10.1101/gr.213611.116.