# Functionality of Non-coding DNA Prediction

**Eshed Gal**
Tel Aviv University
eshedgal@gmail.com

**Yuval Ramot**
Tel Aviv University
yuvalramot@gmail.com

**Lihu Zur**
Tel Aviv University
lihu.zur@gmail.com

## Abstract

Bio-informatics is a research field which involves using computer science and computational tools for medical and biological purposes. Using computer science and specifically machine learning techniques for such research offers a significant power and allows conducting research in ways and scales that are not possible otherwise. Important area of research is the study of the DNA. Our research focuses on non-coding parts of the DNA, and specifically in its role of projecting autism appearance in children. While the coding part has been a target of a lot of research over the last few years, nowadays it is understood that some of the answers to questions such as disease appearances can't be answered through additional research of this coding part, and deeper analysis of the non-coding part is needed. Our aim is to recreate and make accessible former research models of this topic, and use those models for further analysis. We use machine learning and deep learning models and techniques to achieve this goal, and give the research lab the ability to continue it's research using those tools.

## 1 Introduction

Studying the non-coding part of the DNA has important results in various fields. In the case of autism appearance in children, for example, it was shown that in 75 percents of the subjects, the coding part appeared to be normal. Hence, the reason these diseases appear may lie somewhere else in the DNA. This is the non-coding part of it, and it holds about 97 percents of the DNA – which means it is much more complicated to investigate. The lab's research goal is to try and find functional sequences that might suggest a disease among this non-coding part.

Previous work has managed to create some classification of the relation between DNA, cells and their significance in disease appearance. Our goal is to recreate this work and make it available for the lab, and specifically runnable on the lab's resourses. This will make further research, such as with different cell types or data, easily possible for the research group. In our work, we recreate the Deepsea model (Zhou and Troyanskaya, 2015). The model uses deep learning techniques and convolutional network to classify specific functionalities given a DNA sequence. With those results in hand, we use another model, TF-MoDisco (Novakovsky et al., 2023), to indicate the importance of a change in a single DNA sequence. This paper presents the outline of creating this pipeline for future use. As stated above, the lab's use of this pipeline is expected to be significant. In the upcoming sections, we will elaborate on each step, creating the untrained Deepsea model and composing it with the TF-MoDisco model. Example of the interpretation process is shown in figure 1.

## 2 Related Work

Two main previous works were conducted for the classifications step. Our main interest, and the model we recreate, is the Deepsea model (Zhou and Troyanskaya, 2015), which uses convolutional neural network. Another work with similar purpose is the SEI model (Chen et al., 2022). After the classifications step, there are several ways to indicate the importance of change in a single DNA sequence. Here, we use the TF-MoDisco model (Novakovsky et al., 2023). The connection between the Deepsea model and the TF-MoDisco model has an intermediate step, using the Deeplift algorithm (Shrikumar et al., 2019). Another approach for this step is using Captum library (Kokhlikyan et al., 2020).

## 3 Deepsea Model

In this section, we will elaborate on the recreation of the Deepsea model (Zhou and Troyanskaya, 2015). As this model was presented in 2015, its

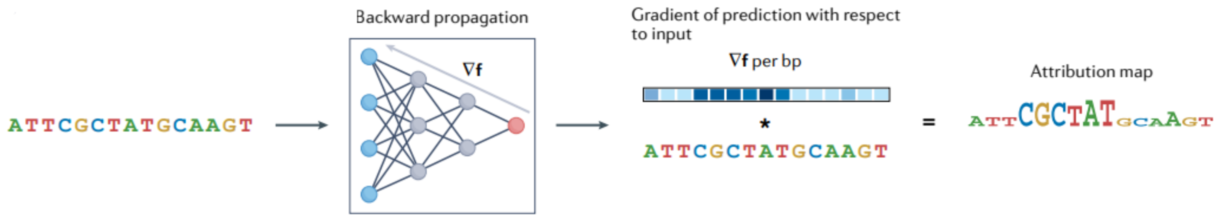

Figure 1: Abstraction of the interpretation process

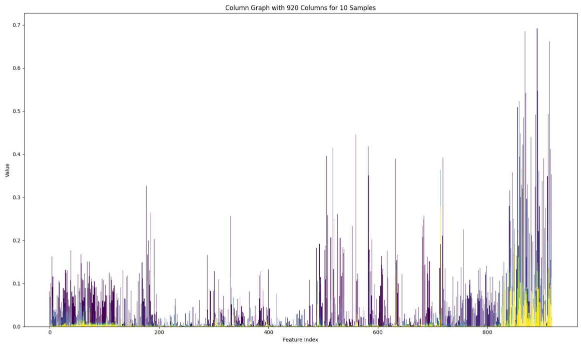

Figure 2: Visualization example of the distributions results for Deepsea over 10 sequences. Each sequence is in different color, and the plot shows values as function of feature index

original shape was using python 2 and several complex packages, such as lua. Our first aim was to use this original version, but it was proved to be very technically difficult to run python 2 based model on the university servers, especially since python doesn't support backward compatibility.

Our solution was using another approach, which includes adaptation of the model to python 3. Using Kipoi (Žiga Avsec et al., 2018), we were able to recreate the model architecture on a pytorch base. We recreated the model architecture in the untrained form, allowing the lab to retrain the model or to finetune it as they wish. Example of visualization of the Deepsea results can be found in figure 2. The model architecture includes 3 iterations of conv-2D, RELU and max pooling, followed by fully connected, RELU, another fully connected and sigmoid layers. Presentation of this model in the untrained form has been achieved as our first milestone of this project.

## 4 Scores Importance Step

After the classification step, the next thing in the pipeline is the model that obtains the importance and indications of the Deepsea's scores. The model for this step is the TF-MoDisco (Novakovsky et al., 2023), and using it requires an intermediate step of an algorithm that analyses the scores returned by Deepsea.

### 4.1 Deeplift

Our first try was using TF-Modisco after the Deeplift algorithm (Shrikumar et al., 2019). This approach failed as there were significant technical difficulties with this method. Specifically, as described above, and by the lab's agreement, our new Deepsea model is built using pytorch. On the other hand, the Deeplift algorithm uses keras packages. Since the connection between pytorch and keras is technically challenging, we chose to find different approach, which we found using Captum.

### 4.2 TF-Modisco Using Captum

Our second and successful try was using Captum (Kokhlikyan et al., 2020) for the intermediate step algorithm. Captum is a model interpretability and understanding library for pytorch. Using its properties and specifically the IntegratedGradients method we were able to obtain the scores needed for the TF-MoDisco model. TF-MoDisco recieves the following inputs: 1. An N x L x 4 array of one-hot encoded genomic sequences, where N is the number of sequences and L is the sequence length (the 4 bases are in A, C, G, T order); this denotes the identity of the sequence. 2. A parallel N x L x 4 array of contribution scores; each position contains the importance of the base specified in the corresponding one-hot encoded sequence (i.e. each base position should have at most one nonzero entry out of the 4, which measures importance at the base in the sequence). This data is the results of the IntegratedGradients method. 3. An optional input of hypothetical contribution scores, of same dimensions. The results of TF-MoDisco allow understanding the scores returned by Deepsea. For a given sequence and feature, the TF-MoDisco shows their connection. Visualization of the input of TF-MoDisco can be found in figures 3 and 4.

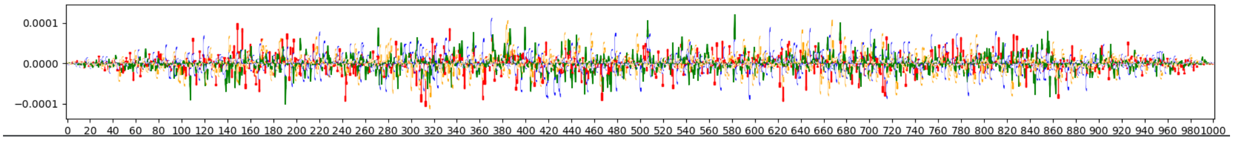

Figure 3: Visualizations example of contribution scores of one sequence over one feature, which is one of TF-MoDisco inputs

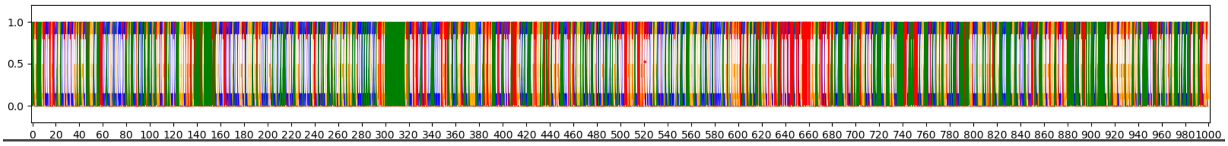

Figure 4: Visualizations example of one hot genomics sequence of one sequence over one feature, which is one of TF-MoDisco inputs

## 5 Results

Our results present the full desired pipeline, using Deepsea, Captum's integrated gradients and TF-MoDisco, as described above. The full code including annotation can be found in the following link: https://github.com/YuvalRM/ML-Workshop/tree/main This full pipeline now stands in hand of the lab for any desired future work.

## 6 Conclusions

This project has been a part of bio-informatics research over the non-coding part of the DNA, and its purpose was to give the lab deep learning tools for their research on various topics, for example the study of autism appearance in children. We presented recreation of the neural network needed to calculate and understand the importance of changes in DNA sequences over different features, by using models based on Deepsea and TF-MoDisco.

## 7 Acknowledgements

We thank Rani Elkon for his guidance and the opportunity to take part in his lab's research and get to know bio-information lab's work. We thank Sapir Shemesh for collaborating with us over the entire project.

## References

Kathleen M. Chen, Aaron K. Wong, Olga G. Troyanskaya, and Jian Zhou. 2022. A sequence-based global map of regulatory activity for deciphering human genetics.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Gherman Novakovsky, Nick Dexter, Maxwell W. Libbrecht, Wyeth W. Wasserman, and Sara Mostafavi. 2023. Obtaining genetics insights from deep learning via explainable artificial intelligence.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2019. Learning important features through propagating activation differences.

Jian Zhou and Olga Troyanskaya. 2015. Predicting effects of noncoding variants with deep learning–based sequence model.

Žiga Avsec, Roman Kreuzhuber, Johnny Israeli, Nancy Xu, Jun Cheng, Avanti Shrikumar, Abhimany Banerjee, Daniel S., Kim Lara Urban, Anshul Kundaje, Oliver Stegle, and Julien Gagneur. 2018. Kipoi: accelerating the community exchange and reuse of predictive models for genomics.