Machine Learning Workshop – Baseline Step

**Functionality of Non-coding DNA Prediction**

Yuval Ramot Lihu Zur Eshed Gal

Background

Our project is a part of the research in Rani Elkon’s lab, under the faculty of medicine.

The research focuses on non-coding parts of the DNA, and specifically in their role of projecting autism appearance in children. The DNA has a part called the coding part, which is about 3% of its length. This part has been a target of a lot of research over the last few years. However, 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.

In the case of autism appearance in children, it was shown that in 75% 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 as mentioned it holds about 97% of the DNA – which means it is much more complicated to investigate. The research goal is to try and find functional sequences that might suggest a disease among this non-coding part.

Project Baseline

Previous work has managed to create some classification of the relation between DNA, cells and their significance in disease appearance. Specifically, there were 2 previous works, DeepSEA which was conducted in 2015, and SEI, which is from 2022.

As to the request of prof. Elkon, our first task is to recreate the results of these works, or at least those of DeepSEA. This will allow his lab to perform further research.

As this work and the data are public, we will read and understand the relevant Git repositories, and use the university servers and computational power to be able to run this learning process and train our own model.

Machine Learning Implementation Techniques

Our Task is a multi-class classification problem; Given a DNA sequence, we will classify it to a specific functionality and many other features like whether it is open in a specific cell type or if it folded, can it be binding with other proteins and more (about 800 different types in deepSEA).

Once we have solved this classification problem, the next step will be to create and train another model (on top of the older one), that indicates the importance of change in a single cell type. With this type of model, we can better understand the non-coding part as a factor in disease appearance.

The learning process is a deep learning module, as we will use the same one deepSEA used, and train their neural network. Further works past this stage might include use of RNNs or other deep learning techniques, as DNA is a sequential data.