Technical Test Wang Lab

**Introduction: with background, principal references, problem statements**

Gene regulation is a fundamental cellular process controlled through complex interactions between regulatory proteins and DNA. Mis-regulation of these regulatory processes can lead to the development of numerous diseases. One technique to uncover the regulatory DNA is DNase1 hypersensitivity sites (DHS) analysis, which utilizes the DNase1 enzyme that is hypersensitive to nucleosome depleted regions in chromatin that are accessible to regulatory proteins.

With the recent rise in sequencing technology, Meulan et al. used Dnase1 hypersensitivity sequencing (Dnase-seq), to map the DHSs from 733 biosamples from 438 cell and tissue types and states. The expansive analysis uncovered a wide range of DHSs, that vary depending on the cell state and tissue of origin. To further study the disease and phenotype context of the regulatory maps, new computational biology methods are being developed.

In recent years, the focus has shifted to Large Language Models (LLMs) pre-trained on DNA sequences. These models may offer an alternative to current bioinformatic methods that rely heavily on experimental annotation, feature engineering and introduce bias. LLMs instead could potentially predict DHS disease phenotype solely based on the sequence of nucleotides. I propose to utilize nucleotide transformer, a large language model pre-trained on the human reference genome, to predict the association of DHS DNA sequences to their cell type.

**Methods**

Dataset

The dataset of known DHSs was taken from Meuleman et al. 2020 paper: DHS\_Index\_and\_Vocabulary\_metadata.tsv. Selected were 4 cell types: type: GM12878 (replicate: ENCLB441ZZZ), HepG2 (replicate: ENCLB029COU), K562 (replicate: ENCLB843GMH) and hESCT0 (replicate: ENCLB449ZZZ). For each cell type, 11968 exclusive/unique peaks were extracted for a total of 47872 peaks. The nucleotide sequences for each associated DHSs were then pulled from the human reference genome hg38. Dataset was split into: Training (60%), Validation (20%) and Test (20%).

Fine-tuning Nucleotide Transformer

I employed a 500M parameters Nucleotide Transformer model pre-trained on the human reference genome. Using the Huggingface Trainer API, I fine-tuned the pre-trained model by utilizing the DNA sequences of the 4 different cell types on the training and validation set. The learning rate was set to 1e-5. The model was fine-tuned for maximum 5000 steps, with early stopping after patience 3. The training batch size was set to 16 for fine-tuning. Due to limited availability of the GPUs on my HPC, I utilized one L4 GPU in Google Colab.

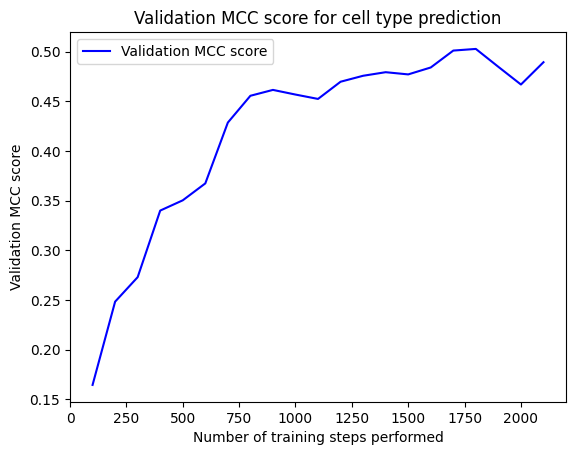
Embedding Extraction and unsupervised prediction of cell type sequences

After fine-tuning the model on the selected cell types, I created embeddings of the sequences in the validation set. Each embedding of sequence was 1280 in length. Uniform Manifold Approximation and Projections (UMAP) and Principal component analysis (PCA) were performed on the 4 different cell types using Python version 3.10.12. Maximum Explained Variance was extracted as a measure of context learning.

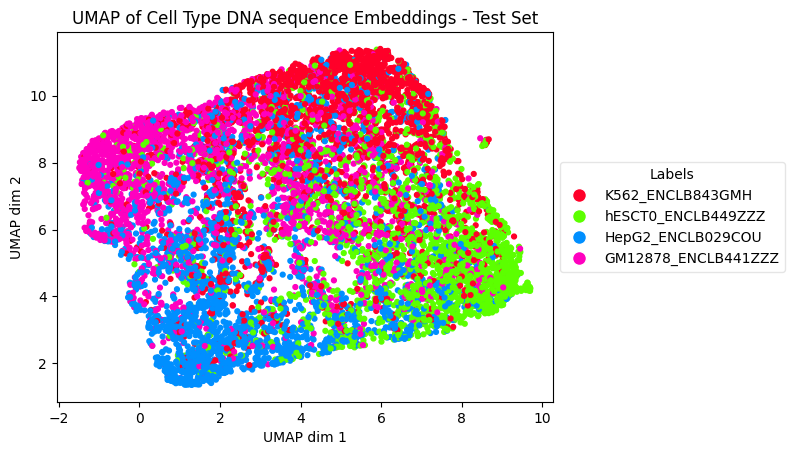
**Results**

To investigate the use of a nucleotide Large Language Model to distinguish between cell types of exclusive DHS peaks, I used the nucleotide transformer model trained on the human reference genome. I selected four distinct cell types: primitive / embryonic stem cell line (hESCT0), Myeloid / erythroid cancer cell line (K562), Hepatocellular carcinoma cell line (HepG2), and a blood cell line (GM12878). After fine-tuning the pre-trained nucleotide transformer model, I evaluated the model accuracy on the validation set, which had 62% prediction accuracy. Additionally, I looked at the MCC score (Fig. 1), and F1 score (Suppl. Fig. 1). Both metrics show an upward trend, indicating an improvement in precision and recall in model’s prediction. However, the model seems to plateau at the ~60% accuracy.

**Figure 1.** MCC Validation Score for cell type prediction during fine-tuning of 500M\_human\_ref model



I further wanted to investigate the extent to which the model learned the cell type differences. I therefor extracted the embeddings of the held-out test dataset using the fine-tuned model. I then applied UMAP (Fig. 2) and PCA (Suppl. Fig. 2), to see if the embeddings hold any predictive power. Both the PCA and UMAP analysis demonstrate that each of the cell type DNA sequences contain distinct features but there is still significant overlap between cell types.



**Figure 2.** Uniform Manifold Approximation and Projection (UMAP) of test set DNA sequence embeddings for 4 cell types: primitive / embryonic stem cell line (hESCT0), Myeloid / erythroid cancer cell line (K562), Hepatocellular carcinoma cell line (HepG2), and a blood cell line (GM12878)

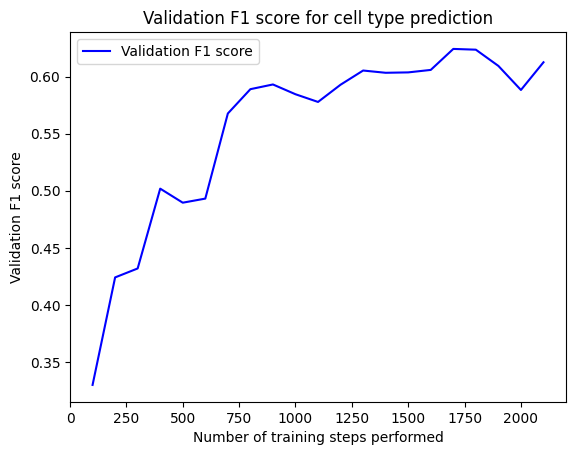
**Discussion**

I fine-tuned a nucleotide transformer on DHSs sequences from 4 distinct cell lines. The goal was to discover if the sequences of the cell types could be distinguished solely based on the nucleotide sequence. The fine-tune model was able to correctly predict some DHS sequence’s cell type, but not with high accuracy. Other researchers have previously fine-tuned nucleotide transformers on genome regulatory sequences (reference), but their approach predicted TATA promoter vs. non- TATA promoters (Reference) or enhancers vs. non-enhancer regions. A major challenge in my approach is that I am not trying to predict a positive and negative labels (binary classification), but 4 different labels. In addition, although the DHS sequences used are mapped to distinct genomic regions, regulatory regions are highly evolutionary conserved and therefor share similar protein recognition motifs, such as the TATA repeat region. In addition, regulatory regions are influenced by various cellular factors, including chromatin state, transcription factor binding, and epigenetic modifications. Solely relying on nucleotide sequence may not capture the full regulatory process, and therefore not lead to high prediction accuracy.

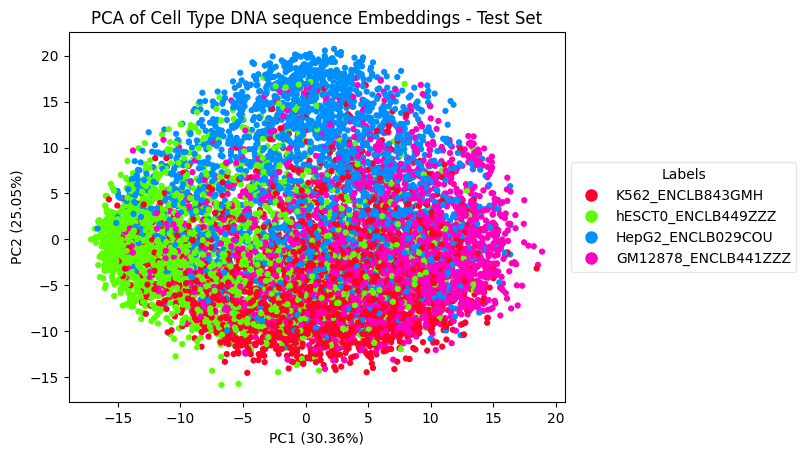
To further improve the predictions, a model with greater parameters could be used, such as the **nucleotide-transformer-2.5b-multi-species. The added model complexity, trained on diverse species genome could increase the models understanding of more subtle distinction in the sequences. Furthermore, I currently extract only the model embeddings from the last layer for the PCA and UMAP. Additionally, extracting embeddings from other layers of the model, may help model to distinguish more nuanced cell type .**

In summary, I explore the use of LLM pre-trained on the human reference genome as potential approach for cell type classification. The low prediction accuracy raises the need to further study the models and assess potential improvements in model architecture, and interpretability.

**Supplementary Figure 1.** F1 Validation Score for cell type prediction during fine-tuning of 500M\_human\_ref model



**Supplementary Figure 2.** Principal Component Analysis (PCA) of held-out test set DNA sequence embeddings for 4 cell types: primitive / embryonic stem cell line (hESCT0), Myeloid / erythroid cancer cell line (K562), Hepatocellular carcinoma cell line (HepG2), and a blood cell line (GM12878)



**References**