Transformer Fine-Tuning for Regulatory DNA: Classifying Functional Elements and Scoring Variant Effects

Angel Morenu
M.S. Applied Data Science
University of Florida
Email: angel.morenu@ufl.edu

***Abstract*—This project will evaluate transformer-based language models for DNA sequence analysis on regulatory genomics tasks. The aims are: (i) to classify functional elements such as promoters, enhancers, and DNase-accessible regions directly from sequence and (ii) to prioritize the regulatory impact of noncoding variants. I will fine-tune DNABERT-2 and Nucleotide Transformer models and benchmark them against convolutional neural network (CNN) baselines (DeepSEA, Basset, Basenji) and linear probes on frozen embeddings. Public datasets from ENCODE, Roadmap Epigenomics, and the DeepSEA training bundle will provide standardized annotations for DNase, histone marks, and TF binding. Primary metrics will be AUROC and average precision (PR-AUC); secondary metrics will include compute cost (runtime, GPU memory) and cross-cell-type generalization. For variant effect prediction, in-silico mutagenesis will be used, with score distributions validated against published DeepSEA benchmarks. The purpose of this study is to determine whether transformer-based models, as opposed to conventional CNN techniques, will be more successful in capturing higher-order dependencies in DNA sequences.**

I. Plan of Action

A. What I Will Implement

1) Data preprocessing pipeline for hg19/hg38 sequences: extract ±1kb windows, build labeled train/validation/test splits from ENCODE/DeepSEA annotations.
2) Baseline models:
   - CNN baseline (Basset/Basenji minimal config).
   - Linear probe on frozen transformer embeddings.
3) Transformer models: fine-tune DNABERT-2 and Nucleotide Transformer with variable k-mer/BPE tokenization and context length.
4) Variant effect prediction: in-silico saturation mutagenesis on known regulatory loci.
5) Evaluation pipeline: AUROC, PR-AUC, bootstrap confidence intervals, runtime profiling, cross-cell-type transfer experiments.

B. Methods to Compare and Sources

- DNABERT-2 (HuggingFace) (HuggingFace) https://huggingface.co/zhihan1996/DNABERT-2-117M
- Nucleotide Transformer (HuggingFace v1/v2) https://huggingface.co/InstaDeepAI/nucleotide-transformer-v2-50m-multi-species
- **Basenji (Calico Labs):** https://github.com/calico/basenji
- **Basset:** https://github.com/davek44/Basset
- **DeepSEA:** https://deepsea.princeton.edu/help/ (portal: https://deepsea.princeton.edu/) *Optional ready-to-use models:* Kipoi https://kipoi.org/models/DeepSEA/predict/, https://kipoi.org/models/DeepSEA/variantEffects/

C. Datasets and Sources

- DeepSEA training bundle: http://deepsea.princeton.edu/help/
- ENCODE Project: https://www.encodeproject.org/

- Roadmap Epigenomics Data: accessible via the NIH Roadmap Web Portal (WashU): https://egg2.wustl.edu/roadmap/web_portal/, also mirrored via AWS Open Data (https://registry.opendata.aws/roadmapepigenomics) and GEO (NCBI) listings.

D. Experiments and Measurements

- Multi-label classification (promoters, enhancers, DNase, TF binding, histone marks).
- Metrics: AUROC, PR-AUC, runtime, GPU memory, cross-cell-type generalization accuracy.
- Variant effect prediction: mutagenesis scores compared with DeepSEA variant scoring benchmarks, including score correlation and enrichment analysis.

E. Feasibility Considerations

Large-scale transformer training can be computationally intensive. This project will use pre-trained checkpoints and fine-tune with smaller sequence windows (1kb) on available university HPC resources or Colab Pro environments. Baselines (Basset/Basenji) are lightweight, ensuring comparisons remain feasible.

II. Preliminary Reading List

1) Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, 2021.
2) N. Nguyen, D. Tran, et al., "DNABERT-2: Efficient foundation model for DNA language," *bioRxiv*, 2023.
3) F. Dalla-Torre et al., "Nucleotide Transformer: Building an LLM for genomics," *arXiv preprint arXiv:2306.15006*, 2023.
4) J. Zhou et al., "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature Methods*, vol. 12, no. 10, pp. 931–934, 2015.
5) D. R. Kelley, J. Snoek, and J. L. Rinn, "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks," *Genome Research*, vol. 26, no. 7, pp. 990–999, 2016.
6) Z. Avsec et al., "Effective gene expression prediction from sequence by integrating long-range interactions," *Nature Methods*, vol. 18, pp. 1196–1203, 2021.