Project Proposal - Can a Self-Attention Model Capture Biological Context of Sequences to Classify Enhancer Regions?

ICS 675

William Little (30286169) October 25, 2025

Introduction & Problem Statement

This project lies in the general area of gene regulation, with a focus on the identifying enhancer regions in the human genome. Enhancers are short, non-coding DNA sequences that regulate gene expression by interacting with promoters and transcription factors. Unlike coding regions, enhancers do not follow a simple positional grammar and they can have regulatory effects from thousands of base pairs away from the genes they control. This complexity leads to enhancers being more difficult to identify computationally than protein-coding genes, which typically contain recognizable open reading frames.

The question being investigated is: how can a deep learning model identify enhancer sequences from random background DNA purely from their sequence composition? From a biological perspective, answering this question is essential because enhancers control cell type–specific expression programs and are frequently disrupted in human disease. From a computational perspective, this task offers a way to test whether modern deep learning methods, specifically self-attention models in this research, are capable of extracting biologically meaningful sequence patterns. Success in this task provides evidence that lightweight attention models can be effective in genomics.

Hypothesis

If a lightweight model composed of a k-mer encoder, self-attention block, and classifier head is trained on the human_enhancers_ensembl dataset from Genomic Benchmarks [1], then it will achieve classification accuracy above random chance, because the attention mechanism can learn to capture both local sequence motifs and broader contextual dependencies that characterize enhancer regions.

Dataset

This project will use a publicly available dataset curated from the Genomic benchmarks paper [1]. A collection of datasets were curated for the purpose of benchmarking deep learning models for classification of genomic sequences. The specific dataset is called the human_enhancers_ensembl, created with data downloaded from Ensembl and consists of 154,640 base pair reads with an equal split of enhancer and noise labels. The sequence length stats are $\mu_L = 270$ and $\sigma = 123$. Positive sequences are from The FANTOM5 project, but accessed via Ensembl. Negative sequences are

randomly generated from Human genome GRCh38 to match lengths of positive sequences and to not overlap them.

Experiment Plan

A deep learning architecture will be designed for this classification task using a basic encoder to embed DNA k-mers, a self-attention model [3], and a classifier head to make the enhancer vs noise prediction. The purpose of the encoder is to learn an abstract yet meaningful high-dimensional vector representation of k-mers, such that close embeddings in the vector space have similar biological meaning. A self-attention block is then applied to the sequence in its embedding representation. Attention is a powerful deep learning technique that, for the purpose of this use case, computes the 'attention' between every k-mer in the sequence. The attention block outputs a new representation of the sequence that contains critical information of the relationships between every k-mer that the model learns is important in identifying enhancer regions. A pooling method is then used on the output sequence, such as using the mean of the sequence, to reduce its dimensionality to a vector. This vector is then applied to a simple feed forward network to perform the binary classification. A visual representation of this architecture is shown in Figure .

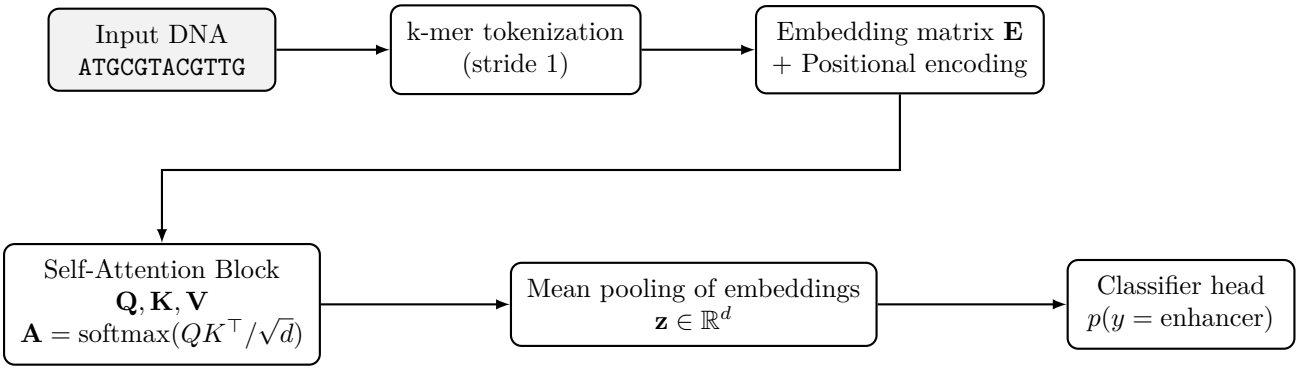

Figure 1: Two-row architecture: DNA → k-mer tokenization → embeddings (top row), then attention, pooling, classifier (bottom row).

The dataset from Genomic benchmarks provides a balanced 50/50 train–test split, with 77,320 sequences available for training. Given resource and time constraints, the scope of this work is limited to relatively lightweight attention-based architecture rather than a large foundational model such as DNABERT (96M parameters) [2]. The task at hand is a focused downstream binary classification problem for classifying enhancer regions with sequences averaging only 270 base pairs. The hypothesis believes that a compact, low-parameter model is sufficient to capture the relevant sequence patterns without the overhead of pretraining a genome-scale transformer.

Once the model has undergone the exhaustive training process, it will be used to perform inference on the test set to validate its performance on unseen data. The output of this testing procedure will provide a probability of model confidence for each test input sequence that it is

an enhancer region. Since the training data is a 50/50 split of enhancer regions and random non-overlapping regions of GRCh38 with equal length, if the model exhibits greater than 50% accuracy, it is considered a success. This indicates the model has learned biologically meaningful representations and relationships in the training sequences and is able to apply that knowledge to unseen sequences.

Besides the accuracy metric, there are several other classification metrics that provide meaningful insight into how the model behaves in unseen data. The Receiver-Operator-Characteristic (ROC) curve visually displays how the model performance reacts to varying the threshold parameter (i.e. >50% threshold probability = enhancer region), where the area under the curve (AUC) is a critical performance metric. A value close to 1 indicates perfect classification ability. Precision is a metric that describes what percent of the positive predictions made by the model were correct. Similarly, recall describes the model's ability to find all the relevant cases, calculated as the number of true positives over the true positives plus false negatives.

In conclusion, the research objective is to prove the stated hypothesis by training a simple single-head attention classification model to classify enhancer sequences better than random chance. A test accuracy score of <50% would refute the hypothesis, indicating small-parameter self-attention models are not useful for learning biologically relevant relationships and context in 100-300 bp length sequences.

Timeline

The project will follow a structured week-by-week plan to ensure feasibility of scope and progress. Table 1 outlines the major milestones. The focus is on implementing the core attention-based model, applying it to the human_enhancers_ensembl dataset, and evaluating results with appropriate performance metrics.

Week 9 (Oct 23)	Finalize and submit project proposal.
Weeks 10–11	Implement and debug the core attention model on small synthetic test cases. Download, tokenize, and format the dataset for training.
Week 12	Apply the model to the real dataset; develop main analysis and training scripts. Begin initial hyperparameter tuning.
Week 13 (Nov 21)	Mid-term checkpoint: demonstrate a working model on a subset of the data with preliminary results.
Weeks 14–15	Scale training to the full dataset. Run cross-validation experiments, generate plots (ROC, precision/recall, accuracy). Begin drafting the final report.
Finals Week (Dec 17)	Submit final codebase, results, and written report.

Table 1: Planned timeline for project execution.

References

[1] Katarína Grešová, Vojtěch Martinek, Daniel Čechák, et al. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25, 2023.

[2] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017.