Explainable AI for Genomic Tasks with Genome Language Model

CAP 5610 - Machine Learning

Spring 2025

Instructor - **Mengxin Zheng**

Group Members

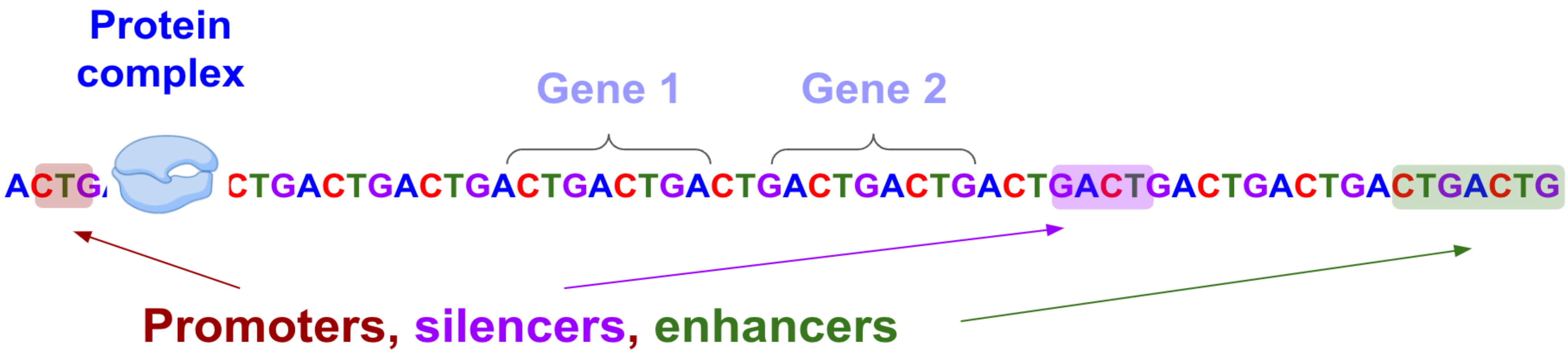
Sourav Saha
Graduate Research Assistant
Dept of Computer Science
Deep Learning, Bioinformatics, Language Models

Nowfel Mashnoor
Graduate Research Assistant
Dept of Computer Engineering
Hardware Security Systems

Abdur Rouf
Graduate Research Assistant
Dept of Computer Engineering
Deep Queue Networks, Speculative SDN

Sagor Biswas
ORCGS Doctoral Fellow
Dept of Computer Engineering
AI Security, ViT, LoRA

Lets Talk DNA



Promoters, silencers, enhancers

(DNA motifs)

Can be very far from binding site, altering chance of protein binding

Motivation

- **Advanced Genome Language Models**: Transformer-based models such as DNABERT, DNABERT2, Nucleotide Transformers, and HyenaDNA have demonstrated strong performance on various genome-specific classification tasks.

- **Knowledge Gap**: Despite their success, the depth of these models' understanding of genomic concepts remains unclear.

- **Explainability Focus**: This project aims to investigate how well these models can explain their predictions by applying them to a downstream interpretability task.

Task Description

- **Objective**: Use advanced genomic language models (e.g., DNABERT, HyenaDNA) for promoter detection.

- **Focus**: Investigate whether these models truly leverage the presence of TATA boxes for classification.

- **Hypothesis**: TATA boxes are key signals that significantly influence the model’s decision-making in identifying promoters.

- **Explainability Goal**: Demonstrate how attention or feature-attribution methods can reveal the role of TATA box motifs in the models’ predictions.

Dataset

Category	Details
Dataset Used	Genome Understanding Evaluation (GUE) for human promoter detection
Sequence Length	300 Nucleotides
Dataset Split	**Training Set**: 47.4k samples **Validation Set**: 5.2k samples **Test Set**: 5.2k samples
Class Distribution	Balanced (Positive:Negative = 50:50)

Models Used

Model	Purpose
DNABERT	Transformer-based model for DNA sequence analysis
DNABERT2	Improved version of DNABERT with enhanced performance
Nucleotide Transformers (NT)	General-purpose nucleotide sequence transformer
HyenaDNA	A novel architecture designed for genomic data

Tools for Explainability in ML

- **LIME (Local Interpretable Model-Agnostic Explanations)**

  Works with any black-box model and provides human-interpretable explanations. Used on text classification(NLP), image classification, and Tabular Data (fraud detection, medical diagnosis)

- **SHAP (SHapley Additive Explanations)**

  Works with various model types (tree-based, deep learning, linear models) and provides both local and global explanations)

- **ANCHOR (High-Precision Model-Agnostic Explanations)**

  More stable model-agnostic tool that provides high-confidence explanations. Used for Text classification, Image classification, Fraud detection, and healthcare)

- **Captum (PyTorch-Based Explainability Library)**

  Captum is a PyTorch-native explainability library that provides various interpretability techniques

Project Schedules

Week	Task	Expected Outcome
1st	Dataset Collection, Preprocessing and Exploratory Analysis	A cleaned and formatted dataset with insights into sequence distributions and quality.
2nd	Baseline Model Training (DNABERT, DNABERT2)	Initial training results for DNABERT models, serving as a reference for further improvements.
3rd	Fine tuning pipeline optimization	Optimized training setup with hyperparameter tuning
4th	Fine-Tuning Nucleotide Transformers (NT) & HyenaDNA	Trained NT and HyenaDNA models with comparative performance analysis.
5th	Explainability Analysis and Model Evaluation	Evaluation of models using SHAP, LIME, and Captum for interpretability insights.
6th	Result generation and report preparation	Consolidated findings, performance comparisons, and a well-structured report ready for submission.

Evaluation Plan

Deletion / Insertion Scores

Measuring change in predictions when important features are removed or added

ROAR (RemOve And Retrain)

Measures the importance of features by retraining models without certain features.

Explanation Stability

Consistency of explanations for similar inputs.

User Trust and Satisfaction

Questionnaires measuring trustworthiness and usability.

Any Questions?