Explainable AI for Genomic Tasks with Genome Language Model

CAP 5610 - Machine Learning

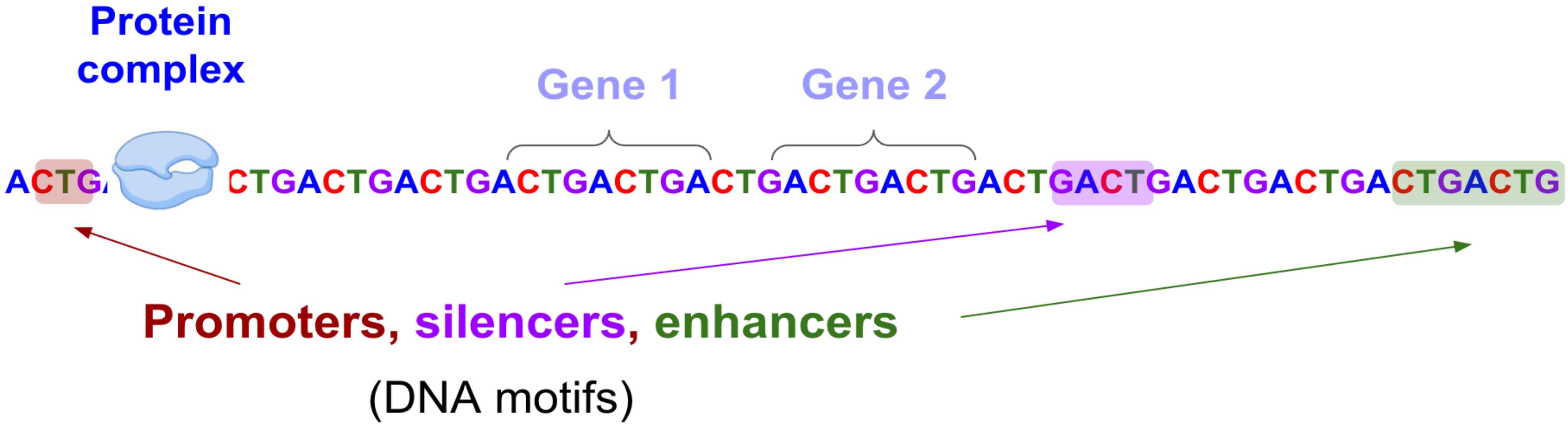
Spring 2025

Instructor - **Mengxin Zheng**

Presented By

Sourav Saha
Abdur Rouf
Nowfel Mashnoor
Sagor Biswas

Lets Talk DNA



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 for Promoter Detection

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

Results

Model	Parameters	Architecture	Tokenizer	Accuracy	MCC	F1 Score	FT Time
DNABERT	1.6 M	Transformers	K-mer	0.969	0.909	0.959	230 m
DNABERT2	117 M	Transformers	BPE	0.932	0.863	0.931	150 m
NT	2.5 B	Transformers	K-mer	0.930	0.853	0.927	120 m
HyenaDNA	1.6 M	Implicit Conv.	Single Nucleotide	0.940	0.862	0.931	35 m

Model Analysis

- **Human-specific pretraining:** Models trained on the human genome excel in species-specific tasks.
- **Multispecies challenge:** Multispecies-pretrained models underperform due to insufficient parameter scaling.
- **Scaling advantage:** Larger models yield better results.
- **Runtime efficiency:** Alternatives to transformer architectures offer faster runtimes.

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)

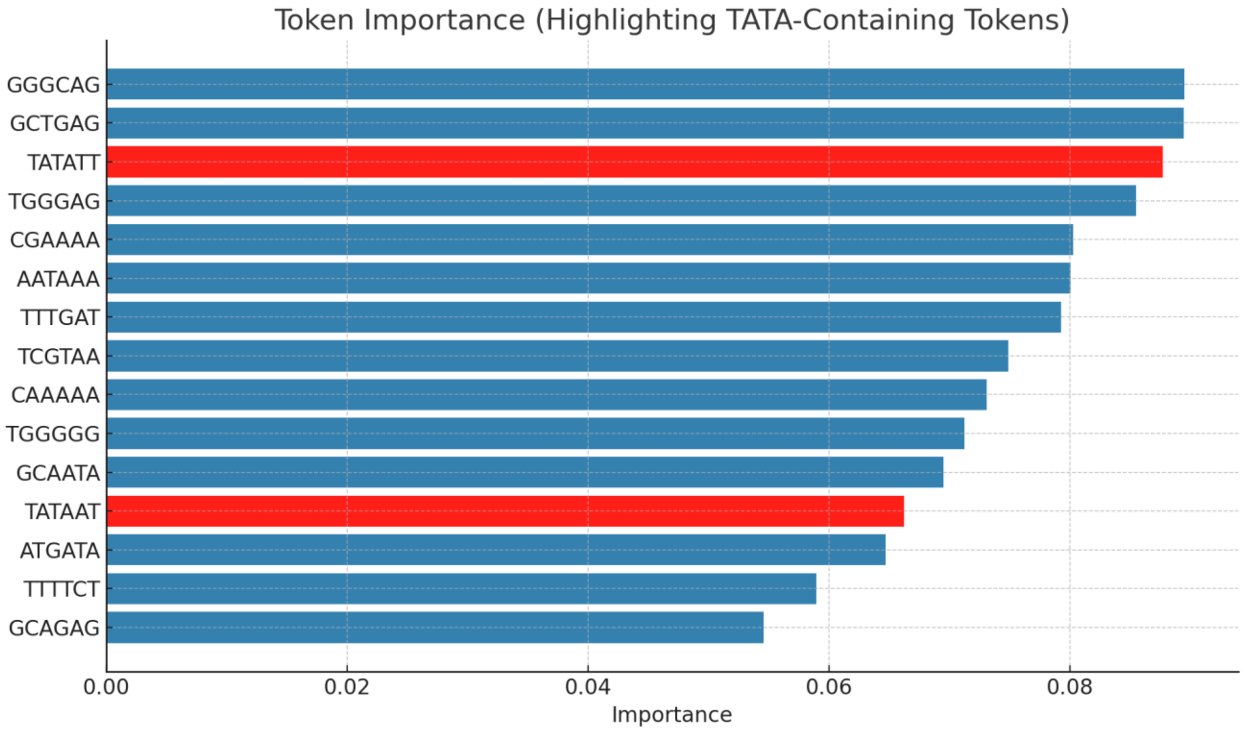
- **SHAP (SHapley Additive Explanations)**

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

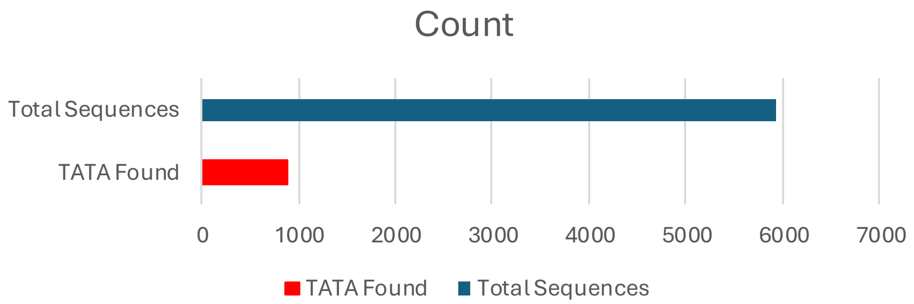
- **ANCHOR (High-Precision Model-Agnostic Explanations)**

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

LIME - Results

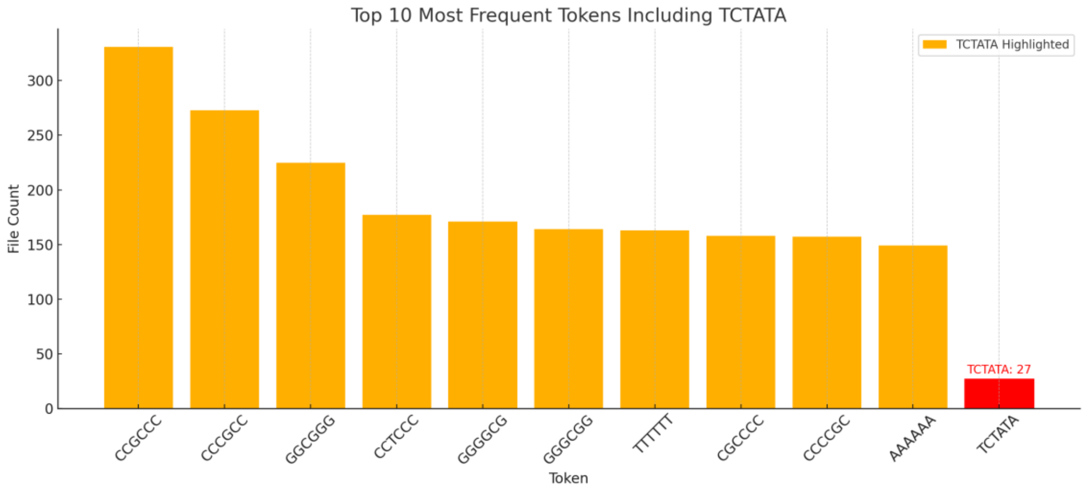


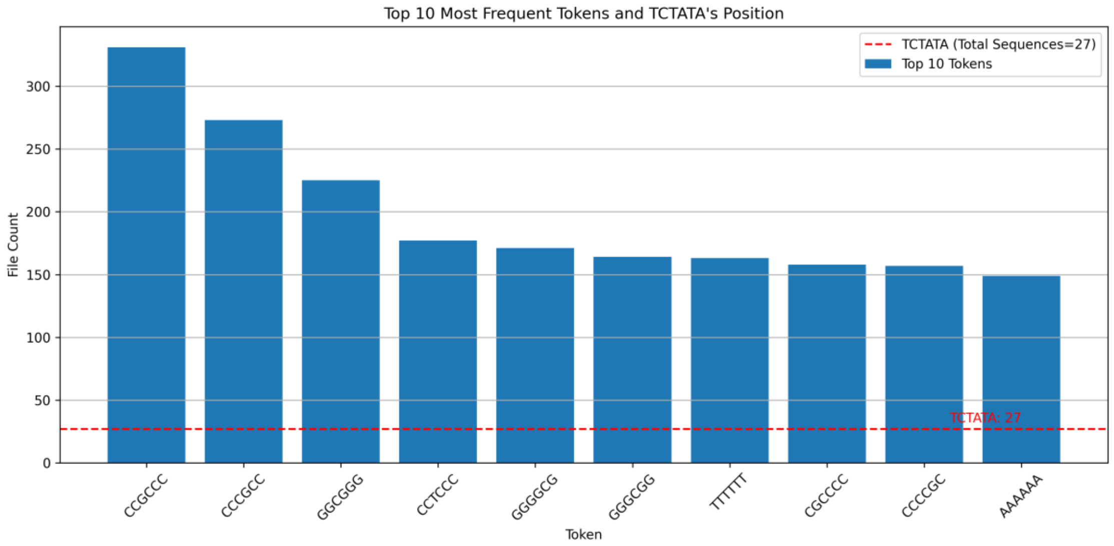
Importance/Label	0	1
Positive	54	99
Negative	639	84



LIME – Results – cont.

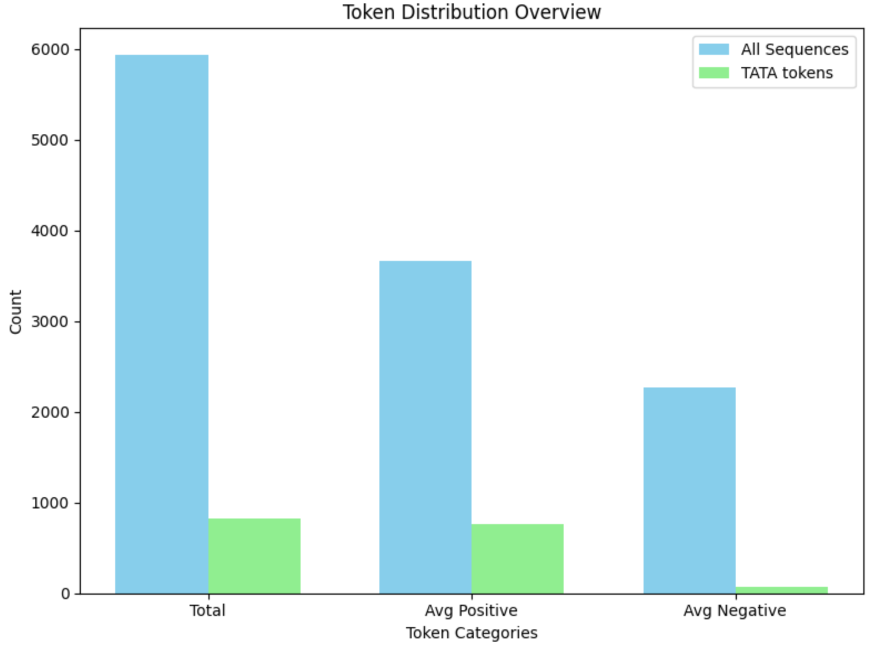
TCTATA, the most common **TATA-containing token**, appears in **27** sequences **(20%)**, whereas the top tokens average around **150** appearances—indicating TCTATA's moderate presence.

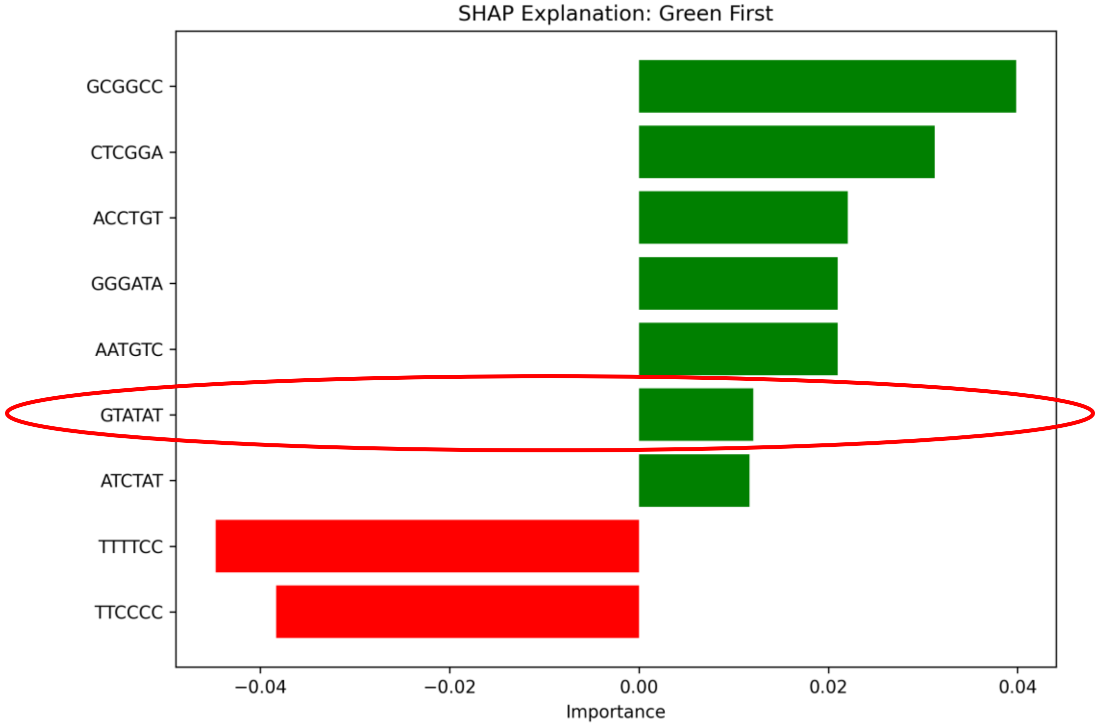




SHAP (SHapley Additive exPlanations)

- Model-agnostic : Can explain any machine learning model
- Unlike LIME, SHAP utilizes Global understanding
- Creates set of features to find importance of each



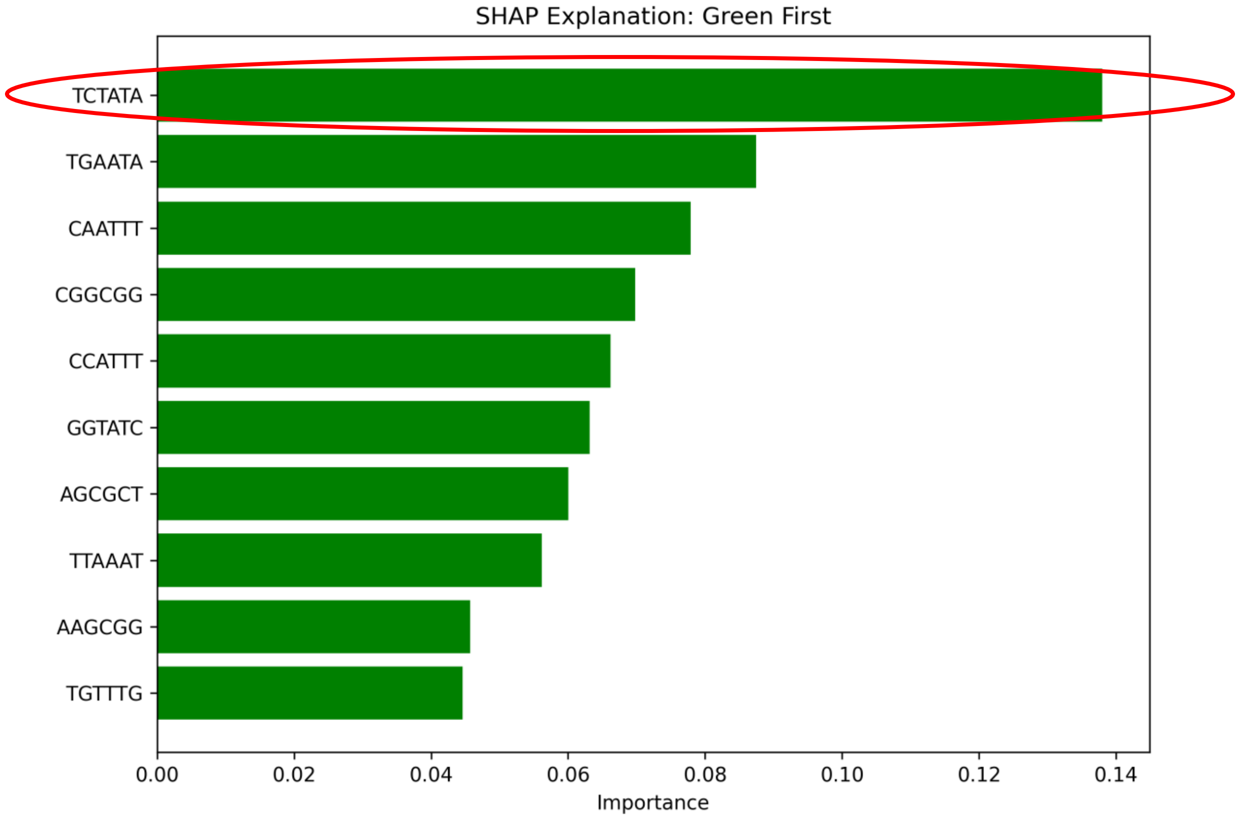


- Total sequence = 5930
- Tokens analyzed per sequence = 10
- Total sequences= 5930
- Total TATA tokens = 828
- Avg Positive tokens = 3662
- Avg Negative tokens =2268
- Positive TATA = 763
- Negative TATA = 65
- Accurate TATA = 277
- Non-accurate TATA=486

SHAP

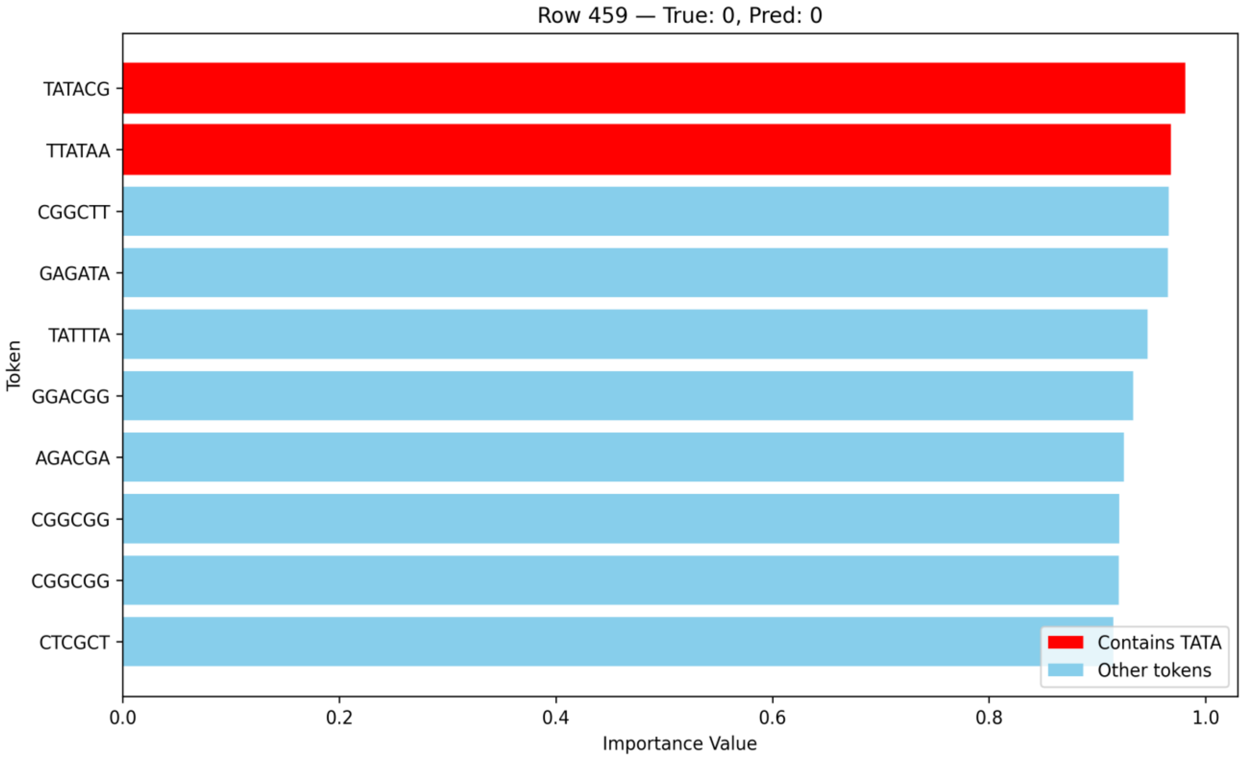
- Top 10 tokens with positive importance

TCTATA	118
CTCCCC	78
GGCGGG	78
GGTGGA	65
GCGGCG	65
GCCTGG	65
TGAATA	65
CCAGAG	65
CCTCCC	65
TATTTT	52

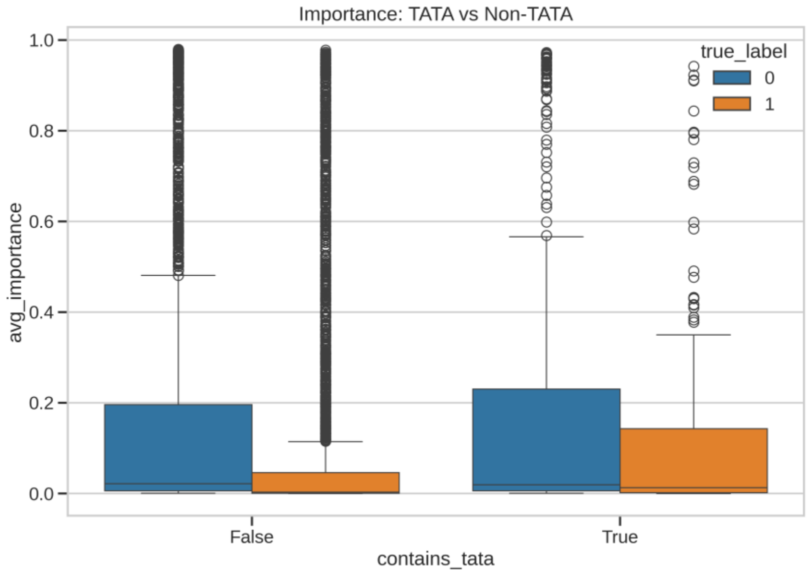


- Model Used : Nucleotide Transformers
- Token size = 6

ANCHOR



Category	Count
Contains TATA	554 (9.34%)
Lable 0 with TATA	407
Label 1 with TATA	147
TATA in Label 0, Negative	0
TATA in Label 0, Positive	407
TATA in Label 1, Negative	0
TATA in Label 1, Positive	147



Limitations

- **Resource**
  - GPU
  - Job Distributor
  - Worker Process (for running the programs independently on 16 GPUs)

- **Time**
  - Analyzing 1 sequence with SHAP tool takes around 500 seconds.
  - Total 5930 sequences take 5930*500 = 824 Hours = 34 Days
  - Had to use parallel processing to complete these tasks within timeline

- **Model Characteristics**
  - Not all the model used for promoter detection are suitable for explainabilty
  - HyenaDNA model cannot be used for our project

Conclusion

- TATA box is a short, specific DNA sequence of about 25-35 base pairs long (TATAAA, or something similar) located at its distance from the transcription initiation site which is part of many promoters.

- It is biologically proved that, "TATA" sequence will be present in the DNA sequence if there is promoter.

- Even ML models, are to some extent dependent on this TATA sequence to detect promoter in DNA

https://github.com/roufunr/explainable-AI-for-genomic-tasks

What??

You Want More??