Deep Sequence Model for Genome Wide Discovery of Coding and Regulatory Element Signatures

Rayhaneh Shabani Nia
University of California, Davis
rshabaninia@ucdavis.edu

Ali Karkehabadi
University of California, Davis
akarkehabadi@ucdavis.edu

Abstract

Deep neural networks achieve strong performance in genomic sequence classification, yet the mechanisms underlying their predictions remain difficult to interpret. This study presents a compact convolutional neural network (CNN) combined with post hoc gradient-based analysis to identify sequence positions that drive discrimination between coding and intergenomic regions. Evaluated on the standardized Demo Coding vs Intergenomic Sequences dataset, the proposed model attains 91.87% validation accuracy while maintaining architectural simplicity. Gradient-based importance analysis reveals nine consistently high-saliency regions across sequences. Clear class-specific separation emerges at positions 20–100 for coding sequences and 150–190 for intergenomic sequences. A mean–variance correlation of $r = 0.530$ indicates stable and discriminative positional signals rather than random fluctuations.

The results demonstrate that lightweight neural architectures can capture biologically meaningful structure without explicit annotations. These findings support the view that predictive accuracy and interpretability can coexist in genomic sequence modeling, enabling hypothesis-driven biological investigation.

1. Introduction

Distinguishing coding sequences from intergenomic regions remains one of the fundamental challenges in computational genomics. Although protein-coding regions comprise only 1–2% of the human genome, their precise identification has broad implications, from genome annotation to disease variant interpretation. Traditional approaches rely on sequence homology, codon usage bias, and probabilistic frameworks such as hidden Markov models. While effective, these methods are constrained by predefined assumptions and may fail to capture context-dependent regulatory structure. Deep learning has substantially reshaped this landscape. Rather than relying on handcrafted features, neural networks learn representations directly from raw sequence data, achieving strong performance in regulatory and gene expression prediction tasks [1]. However, interpretability has not progressed at the same pace. When model predictions inform biological hypotheses or experimental design, it is essential to understand which nucleotides drive decisions and how positional context influences classification. Gradient-based attribution provides a principled mechanism for quantifying input sensitivity [7]. Although early saliency methods were noisy, stabilization techniques such as SmoothGrad and saliency-guided training have improved attribution reliability [3–5]. Their systematic application to coding sequence classification, however, remains limited.

We address this gap with three contributions. First, we demonstrate that gradient-based analysis recovers biologically meaningful patterns, including codon periodicity and localized signals consistent with splice-site and translation initiation structure. Second, we introduce aggregation strategies across samples, nucleotide channels, and class labels to characterize shared positional organization and class-specific signatures. Third, we show that a compact CNN achieves 91.7% accuracy on the Genomic Benchmarks dataset [2] while preserving interpretable structure. Together, these results provide a reproducible framework that connects predictive modeling with biologically grounded interpretation.

2. Related Work

2.1. Deep Learning in Genomics

Deep learning has substantially advanced genomic sequence analysis. Convolutional models have been shown to learn sequence motifs directly from raw DNA [1], enabling regulatory prediction at single-nucleotide resolution [11]. Hybrid architectures and attention-based models further extend this capability to capture longer-range dependencies and distal interactions [6]. Although these models learn informative sequence representations, the specific features driving their predictions remain difficult to interpret.

2.2. Gradient-Based Attribution

To examine model sensitivity, gradient saliency measures the change in prediction with respect to the input:

$$S(x)_c = \left| \frac{\partial f_\theta(x)_c}{\partial x} \right|.$$

This quantity reflects how strongly each input position influences the class score. While computationally efficient, raw gradients may exhibit noise or saturation [8], motivating stabilized variants such as Integrated Gradients [10] and SmoothGrad [9].

2.3. Application to Genomic Sequences

In genomic data, sequences are discrete and position-dependent, requiring careful aggregation of attribution scores. Using the Genomic Benchmarks dataset [2], we adapt gradient-based analysis to compute positional statistics and class-wise differences, enabling structured comparison between coding and intergenomic regions.

3. Methodology

3.1. Dataset and Task Formulation

We study binary classification of coding versus intergenomic DNA sequences using the demo_coding_vs_intergenomic_seqs dataset from Genomic Benchmarks [2]. The dataset contains 100,000 sequences of fixed length $L = 200$, evenly divided between coding transcript regions and non-coding intergenomic regions. A 75–25 train–test split is used.

Each sequence is represented as

$$\mathbf{x} = [x_1, \ldots, x_L], \qquad x_i \in \{A, T, G, C, N\},$$

and a classifier f_θ predicts $y \in \{0, 1\}$ using binary cross-entropy loss.

3.2. Model Architecture

We employ a compact 1D CNN to capture local sequence structure. Nucleotides are mapped to 128-dimensional embeddings, followed by three convolutional layers (kernel size 8; 32, 16, and 4 channels). Each layer includes batch normalization, ReLU activation, and max-pooling (stride 2), enabling hierarchical extraction of motif-level features.

The resulting features (approximately 200 after flattening) are passed through a 512-unit fully connected layer with 0.3 dropout, followed by a two-logit output layer.

3.3. Gradient-Based Importance Analysis

To interpret predictions, gradients are computed with respect to input embeddings:

$$\mathbf{G} = \frac{\partial f_\theta(\mathbf{x})_y}{\partial \mathbf{E}(\mathbf{x})},$$

where $\mathbf{E}(\mathbf{x}) \in \mathbb{R}^{L \times d}$ with $d = 128$.

Position-level importance is defined as

$$I_j = \sum_{k=1}^{d} |G_{j,k}|, \qquad j = 1, \ldots, L,$$

which aggregates sensitivity across embedding dimensions.

3.4. Positional Importance Analysis

Across $N = 1{,}000$ sequences, positional statistics are computed as

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} I_{i,j}, \qquad \sigma_j = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (I_{i,j} - \mu_j)^2}.$$

Positions satisfying

$$\mu_j > \text{percentile}_{75}$$

are considered high-importance regions.

3.5. Class-Specific Importance Comparison

For each class $c \in \{0, 1\}$, average importance vectors $\boldsymbol{\mu}_c$ are computed using 250 sequences. The difference

$$\boldsymbol{\delta} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$$

highlights coding-specific regions ($\delta > 0$) and intergenomic-specific regions ($\delta < 0$).

4. Results

4.1. Model Performance

The proposed SimpleSaliencyCNN achieves strong performance on the coding versus intergenomic classification task. Despite containing only **91,145 parameters**, the model attains a validation accuracy of **91.87%**. As shown in Table 1, performance is comparable to larger sequence models while maintaining substantially lower complexity.

Method	Validation Accuracy
SimpleSaliencyCNN (Ours)	**91.87%**
HyenaDNA	91.31%

Table 1. Validation accuracy comparison.

4.2. Gradient-Based Positional Analysis

Gradient-based importance scores were computed over 1,000 sequences to examine positional sensitivity. The aggregated profile (Figure 2) reveals two consistent high-importance regions around **20–40** and **160–190**.

Class-wise comparison (Figure 3) shows distinct concentration patterns: coding sequences exhibit elevated importance across **20–100**, whereas intergenomic sequences

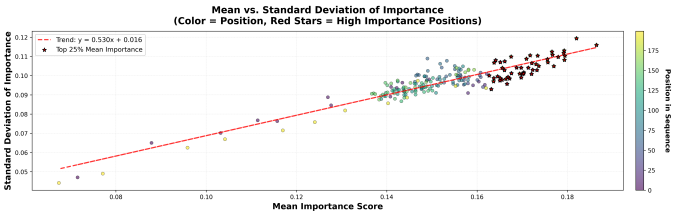

Figure 1. Mean vs. standard deviation of positional importance. Red stars highlight top 25% positions.

emphasize **150–190**. The mean–variance relationship (Figure 1) displays a positive correlation ($r = 0.530$), indicating that positions with higher average influence also exhibit greater variability across samples.

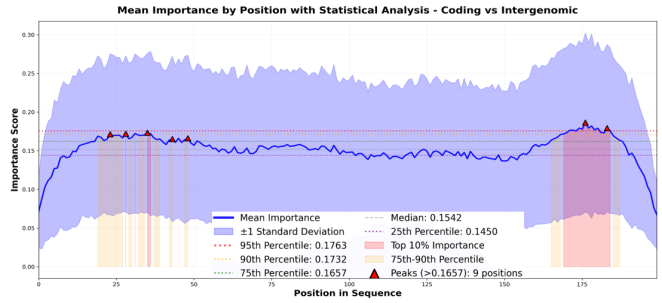

Figure 2. Positional mean importance with confidence bands. Nine positions exceed the 75th percentile threshold.

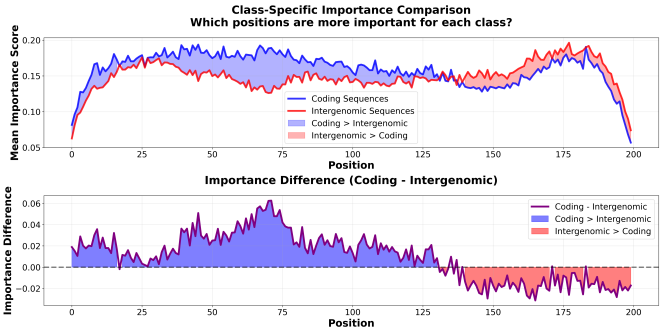

Figure 3. Coding vs. intergenomic importance patterns. Bottom: positional differences.

4.3. Biological Interpretation

The positional patterns align with established structural properties of genomic sequences. Periodic importance in coding regions likely reflects reading-frame organization, whereas localized peaks in intergenomic regions may indicate enrichment of regulatory motifs. The clear separation between classes suggests that the model captures systematic positional and compositional differences rather than isolated sequence artifacts.

5. Conclusion

This study demonstrates that a compact CNN achieves 91.87% validation accuracy on coding versus intergenomic classification while preserving interpretability through gradient-based analysis. Distinct class-specific positional signatures emerge, with coding regions emphasizing positions 20–100 and intergenomic regions 150–190. The positive mean–variance correlation ($r = 0.530$) indicates stable and discriminative positional signals rather than noise. Together, these findings show that lightweight neural architectures can uncover structured biological patterns without sacrificing predictive performance.

References

[1] B. Alipanahi, A. Delong, M. Weirauch, and B. Frey. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. *Nature Biotechnology*, 33:831–838, 2015. 1

[2] K. Grešová, V. Martinek, D. Čechák, P. Šimeček, and P. Alexiou. Genomic benchmarks: A collection of datasets for genomic sequence classification. *bioRxiv*, 2022. 1, 2

[3] A. Ismail, H. Corrada Bravo, and S. Feizi. Improving deep learning interpretability by saliency guided training. In *Advances in Neural Information Processing Systems*, pages 26726–26739, 2021. 1

[4] Ali Karkehabadi, Houman Homayoun, and Avesta Sasan. Smoot: Saliency guided mask optimized online training. In *2024 IEEE 17th Dallas circuits and systems conference (DCAS)*, pages 1–6. IEEE, 2024.

[5] Ali Karkehabadi, Banafsheh Saber Latibari, Houman Homayoun, and Avesta Sasan. Hlgm: A novel methodology for improving model accuracy using saliency-guided high and low gradient masking. In *2024 14th International Conference on Information Science and Technology (ICIST)*, pages 909–917. IEEE, 2024. 1

[6] D. Quang and X. Xie. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic Acids Research*, 44:e107, 2016. 1

[7] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 1

[8] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[9] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 2

[10] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017. 2

[11] J. Zhou and O. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12:931–934, 2015. 1