Pattern Recognition of Core-Promoter DNA Sequences Based on Recurrent Neural Network

Ahmed Oday, Shaimaa Khalid Moufak, Khadija Abbas Sahan, Jafar Mohammed and Mariam Maan Alzaak

Bioinformatics Department, College of Biomedical Informatics, University of Information Technology & Communications, 10001 Baghdad, Iraq

Ahmed.oday@uoitc.edu.iq, shaimaa.khalid-bic@uoitc.edu.iq, dr.Khadija.abbas@uoitc.edu.iq, jafer0028@uoitc.edu.iq, mariamalzaak@gmail.com

Keywords: Core-Promoter, Deep Learning, DNA Sequence, Gene, Pattern Recognition.

Abstract: The interpretation of genomic sequences remains a major challenge in computational biology due to the inability of traditional methods to detect complex, context-dependent regulatory patterns. therefore, detecting the promoter sequences tends to lead to a high false-positive rate. Additionally, to overcome these limitations, we used the Recurrent Neural Network (RNN) framework that autonomously identifies functional genomic elements by modelling long-range dependencies in Deoxyribonucleic Acid (DNA). we proposed the method of combining advanced pattern recognition with experimental validation, outperforming conventional techniques in detecting regulatory motifs while enabling standardized, high-throughput genomic annotation. By bridging computational and molecular biology, this approach provides a powerful solution, including synthetic biology and genome annotation pipelines. Benchmarking results demonstrate that the framework significantly improves detection of non-canonical and weakly conserved regulatory features, which are frequently missed by existing tools. To perform an analysis of the publicly available datasets: GRCh38.p14 on our proposed work, we have analyzed the accuracy is 92.26% in classifying genes and non-genes from the long DNA sequence.

1 INTRODUCTION

Deoxyribonucleic Acid (DNA) is the hereditary material in almost all living organisms, carrying the genetic instructions essential for growth, development, functioning, and reproduction. It has a double-helix structure defined by base-pairing rules: adenine (A) pairs with thymine (T), and cytosine (C) pairs with guanine (G). DNA sequences vary widely in length, ranging from a few hundred base pairs to over three billion in the human genome. These sequences are the subject of extensive bioinformatics analysis, including alignment, motif discovery, variant calling, gene prediction, and functional classification. Many of these computational tasks benefit greatly from the application of deep learning techniques [1]. A common approach to labeling unknown query sequences is to build a database of reference sequences for the property of interest, map the queries to the database using k-mer matching or pairwise-sequence alignment, and finally use the mappings and their scores to assign labels. Alternatively, many important problems in bioinformatics, such as annotating genes and regulatory elements, are addressed by building a statistical model from a relevant database and then using this model to assign labels to queries [2].

The genome refers to the complete sequence of nucleotide bases adenine (A), thymine (T), cytosine (C), and guanine (G) that encodes the hereditary information within an organism's DNA [3]. Not all genomic regions encode for proteins or observable traits. In fact, a substantial portion of the genome is non-coding, meaning it does not directly translate into proteins. This genetic variability introduces challenges in structural genome annotation, which involves identifying functional elements such as genes and regulatory motifs based on DNA sequences, as illustrated in Figure 1. Because individuals within the same species can exhibit significant sequence variation, a genome annotated from one individual may not accurately represent another [4]. Therefore, it is inaccurate to refer to a single, definitive genome for an entire species [5]. A

gene is a coding region of the genome that is organized into structural units with a specific DNA sequence encoding a protein or another functional molecule. In eukaryotic genomes [6], each gene typically begins with non-coding regulatory sequences, followed by an Open Reading Frame (ORF) composed of both coding segments (exons) and non-coding segments (introns), and concludes with additional regulatory regions [7].

The core promoter is a DNA sequence located upstream of a gene's ORF, essential for the initiation of transcription. Virtually every eukaryotic gene possesses at least one promoter, while some are regulated by multiple promoters to allow tissue-specific or condition-specific gene expression. Promoters are broadly categorized into two primary subregions [8]:

Core Promoter. This region spans approximately 50 to 100 base pairs around the transcription start site (TSS). It contains conserved sequence motifs where basal transcription factors (TFs) bind to facilitate the recruitment of RNA polymerase, thereby initiating transcription. Due to its fundamental role, core promoter elements are typically conserved across species.

Proximal Promoter. Located immediately upstream of the core promoter, typically extending up to 250 base pairs from the TSS, the proximal promoter contains binding sites for gene-specific transcription factors such as activators and repressors. These sequences help modulate transcriptional efficiency but are not required to initiate transcription, resulting in higher variability between genes.

Traditional neural networks treat inputs as independent entities, whereas Recurrent Neural Networks (RNNs) are a class of deep learning models that possess internal memory, enabling them to capture sequential dependencies. These dependencies refer to the temporal order of inputs, making RNNs suitable for tasks involving sequential information [7]. RNNs employ a loop that applies the same operation to each element in a series; the current input and the previous computations are both used for the current computation [9]. The ability of RNNs to utilize contextual information is particularly valuable in tasks such as natural language processing, video classification, and speech recognition. For example, in language modeling, understanding the preceding words in a sentence is crucial for predicting the next word. RNNs excel at capturing such dependencies due to their recurrent nature [5]. Computational models should be developed to accurately identify subtle sequence variations, cryptic motifs, and functionally relevant signals or low-affinity regulatory elements that evade detection by conventional rule-based or feature-engineered approaches.

The primary aim of this research is to develop advanced computational methods for the accurate identification of regulatory elements within DNA sequences. The focus is on improving the detection of elusive yet biologically significant features such as weak transcription factor binding sites, cryptic splicing signals, and epigenetic markers that play a pivotal role in gene regulation and disease pathogenesis.

Figure 1 shows a diagram of structural blocks within DNA. In it, Figure 1a represents the positions of eukaryotic genes within the genome. One gene is enlarged to display its inner structure in more detail, and Figure 1b represents the structure of a eukaryotic RNA Polymerase II (Pol II) promoter with a single transcription start site (TSS), as illustrated. The core promoter is enlarged, showing the most commonly found elements along with their consensus sequences. This particular promoter would not be encountered in real biological contexts, as only a subset of these elements is required to initiate transcription. It is therefore used merely as an illustration tool to visualize the approximate location of the individual elements [8].

2 RELATED WORK

The integration of deep learning into genomic analysis represents a transformative shift from traditional approaches that heavily rely on manual feature engineering. Conventional methods are often time-consuming and risk overlooking biologically meaningful signals, especially those that remain poorly characterized or unknown. In contrast, deep learning models, particularly when trained on annotated genomic datasets, can autonomously extract relevant features directly from raw nucleotide sequences. These datasets commonly include known biological elements such as regulatory motifs, splice sites, and transcription factor binding sites, enabling models to learn biologically significant representations with minimal human intervention [1].

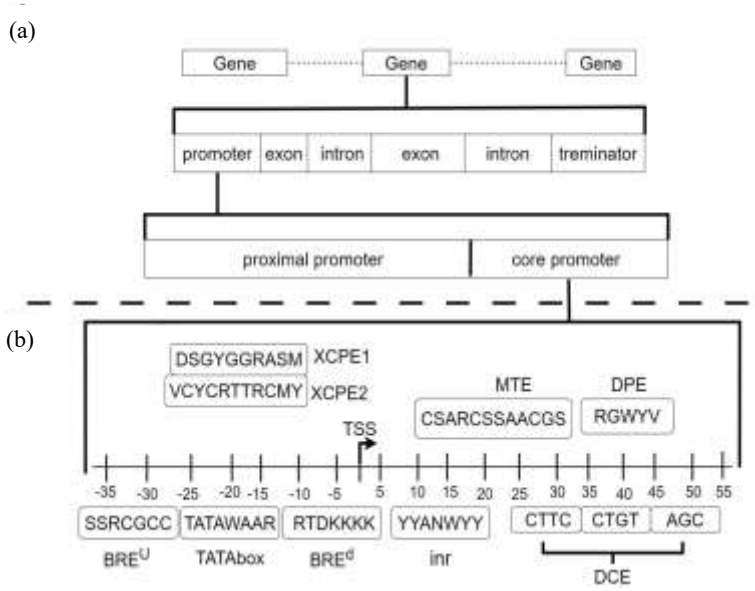

Figure 1: Structural blocks in DNA, a) Positions of eukaryotic genes within the genome, with one gene enlarged to show its internal structure; b) Illustrative RNA Polymerase II promoter with common core elements.

In previous work [3], core promoter elements have been shown to be highly variable, making their detection a complex task that necessitates sophisticated techniques. While ab initio computational techniques are improving, their predictions can still be greatly enhanced by integrating them with experimental data. For example, Zhang [5] employed both linear discriminant analysis and quadratic discriminant analysis to identify human core promoter regions. Furthermore, Ohler et al. [11] demonstrated that promoter predictions improve significantly when starting with accurate transcription start sites determined from a cap-trapped cDNA library with complete 5' ends [1].

Current models often rely on one-hot encoding, which requires uniform input length unless padding or cropping is applied. To address this limitation, a new k-mer encoding strategy has been proposed, which can handle variable-length sequences. Another common issue in genomic analysis, such as splice site prediction, is class imbalance, where negative samples outnumber positive ones. In contrast, promoter annotation remains challenging due to the absence of clear consensus sequences and the difficulty of constructing meaningful negative samples. A novel negative sample construction method has been introduced that retains conserved core promoter elements from positive samples [3].

Despite their diversity, many promoters share conserved elements. Among the most recognized is the TATA box, a DNA sequence typically located 25–30 base pairs upstream of the transcription start site. This element plays a crucial role in positioning the transcription initiation complex [10]. Another conserved feature, often found upstream of the TATA box, is the CAAT box, which helps regulate promoter activity and enhances transcription efficiency [12].

In their work, Nielsen and Voigt [13] explored how deep neural networks can address the sequence-labeling problem, focusing on species-level taxonomic classification of short 16S ribosomal DNA reads. This task presents a suitable benchmark due to its complexity and importance in microbial ecology. Their proposed DNN architecture learns complex sequence-to-label relationships directly, offering advantages over k-mer and alignment-based methods in terms of robustness to sequencing noise and variable read lengths [14]. Separately, a study on RNNs [15] explored how deep recurrent neural network architectures can capture structure within genetic sequences. The authors first confirmed that a character-level RNN can capture non-random parts of DNA by comparing the perplexity after training on a real genome to that from a random nucleotide sequence. Because multiple biological objects can interact with a given sequence, they framed this as a multitask learning problem, empirically showing how a deep network can outperform a baseline model on a significant majority of 919 binary labeling tasks [16]. However, a limitation of simple RNNs is their short-term memory, which restricts their ability to retain information over long sequences. To overcome this, more advanced RNN variants have been developed,

including Long Short-Term Memory (LSTM) [17], bidirectional LSTM, Gated Recurrent Unit (GRU) [18], bidirectional GRU, and Bayesian RNN [12].

A central challenge lies in developing computational frameworks capable of identifying subtle variations, cryptic motifs, and other biologically meaningful signals without relying on manual feature extraction or predefined rules. These elusive patterns, such as weak transcription factor binding sites, non-canonical splicing signals, or low-affinity regulatory elements, can have significant functional implications in gene regulation, protein-DNA interactions, and cellular signaling pathways, despite not conforming to known consensus sequences or scoring low on traditional statistical metrics [19]. Therefore, robust analytical techniques are needed that can learn from raw sequence data, adapt across diverse genomic environments, and reveal meaningful insights beyond the capabilities of conventional approaches.

3 METHODS

3.1 Preprocessing

The genomic database employed in this research was meticulously constructed through a structured data processing pipeline utilizing primary biological data sourced from the NCBI Reference Sequence Database (RefSeq). The human genome assembly in FASTA format served as the reference standard for the study due to its wide adoption and compatibility with bioinformatics tools. Chromosome-scale FASTA files were downloaded directly from NCBI's FTP servers, providing comprehensive DNA sequences for all human chromosomes. These raw sequences formed the foundational dataset used for genomic analysis and model training.

3.2 Computational Genomic Element Detection Framework

The computational framework for genomic element detection maps transcription initiation and termination signals within noisy DNA sequences. For core promoter detection, the BRE motif is first located via "CGCC" detection with upstream "SSR" validation, followed by identification of the TATA box ("TATAWAAR") using standardized shorthand translation, and concluded with Poly-A signal ("AATAAA") verification. All validated element coordinates are returned as a structured dictionary for precise promoter annotation.

For TATA-dependent core promoters, a hierarchical detection method is employed: the TATA box is identified through mandatory "TATA" and "WAAR" component recognition, the Inr motif is verified by confirmation of its central "A" flanked by "YY" upstream and "NWYY" downstream sequences, and the Poly-A signal is validated. Positional data are only returned upon confirmation of strict 5′→3′ element ordering to prevent false positives.

Annotation of TATA-less promoters is achieved through Inr motif ("YYANWYY") validation, followed by detection of the downstream promoter element (DPE) under stringent "RGWYV" contextual constraints (requiring R/G/WYV positional compliance), and Poly-A signal identification. The complete regulatory unit spanning from the Inr start to the Poly-A signal end is output for alternative promoter characterization.

Epigenetic regulatory regions are localized by scanning 200-bp windows to identify GC-rich regions (≥50% GC content and ≥0.6 CpG ratio), with subsequent verification of downstream Poly-A signals ("AATAAA"). Non-overlapping islands are returned with associated epigenetic metrics and Poly-A coordinates.

Finally, precise sequence modifications are enabled through segment excision: user-defined regions are replaced with masking characters ("N") while preserving the original sequence length. Input boundaries are validated, overlaps are prevented, and excised segments are cataloged for downstream analytical integrity. The steps of this computational genomic sequence analysis are illustrated in Figure 2.

Table 1 shows that core promoter genomic elements are critically important because they serve as fundamental regulators of gene expression. Mutations in these highly conserved motifs frequently lead to severe disruptions in transcriptional regulation, which can result in genetic disorders. The presence of these specific sequence patterns provides strong predictive evidence for functional DNA regions.

3.3 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) form a specialized category of deep learning models that are particularly well-suited for analyzing sequential biological data, such as DNA sequences. The genomic sequence analysis in this study was conducted through a carefully designed deep learning architecture that systematically addresses both biological sequence processing and computational challenges.

Initially, DNA sequences are preprocessed by being divided into fixed-length segments of 1000 nucleotides to handle excessive sequence lengths. These segments are then converted into one-hot encoded 3D tensors with dimensions [batch_size, 1000, 4], where the final dimension represents the four nucleotide bases (A, T, C, G). Padding positions in these tensors are automatically identified and ignored through a built-in masking mechanism that utilizes a special value (-1), ensuring only valid nucleotide data is processed.

The sequence processing is performed through a hierarchical bidirectional recurrent neural network architecture. First, a bidirectional LSTM layer containing 256 total hidden units (128 units for forward processing and 128 for backward processing) is applied, with return_sequences=True to maintain the temporal structure. This is followed by a bidirectional GRU layer with 128 total hidden units (64 units in each direction), which provides additional sequence processing capacity. Notably, the masking information is automatically propagated through both recurrent layers.

The processed features are then transformed through a dense processing block consisting of two components: a time-distributed dense layer with 128 units and ReLU activation that operates independently at each sequence position, followed by a dropout layer with a 0.3 rate that randomly deactivates 30% of neurons during training to prevent overfitting. This combination enables robust feature extraction while maintaining positional information. The steps of this architecture are shown in Figure 3.

Finally, sequence annotation predictions are generated by a time-distributed dense output layer that utilizes softmax activation to produce probability distributions at each position, with the number of units corresponding to the annotation classes (two-class). This architecture effectively combines chunk-based sequence processing with advanced neural network components, enabling comprehensive analysis of full-length DNA sequences while maintaining computational efficiency through synchronized parameter updates across all sequence segments during each training epoch. The integration of bidirectional processing and proper sequence masking results in a powerful tool for accurate genomic element identification.

4 RESULTS

4.1 Training Performance

We report the achievable accuracy on the test dataset. For these experiments, the parameters for all models were set to 100 epochs and a batch size of 32, using a softmax activation function. The experiments were conducted on a computational system equipped with an NVIDIA GeForce RTX 4080 GPU (12GB VRAM) and 64GB of system memory. The Adam optimizer was used for model training. The dataset was split into 70% for training, 10% for validation, and 20% for testing. All code and results are available on GitHub [20].

Table 1: Core promoter short and function.

Element	Consensus Sequence	Biological Function
BRE^u	SSRCGCC	Binds TFIIB (transcription factor), ensuring correct RNA polymerase II positioning.
TATA Box	TATAAA	Anchor for TBP (TATA-binding protein); marks eukaryotic promoter regions.
BRE^d	RTDKKKKK	Binds TFIIB downstream of the TATA box, stabilizing the transcription initiation complex.
Inr (Initiator)	YYANWYY	Surrounds the transcription start site (TSS), which is critical for transcription initiation.
DPE	RGWYV	Stabilizes transcription machinery downstream of the TSS (~+28 to +32).

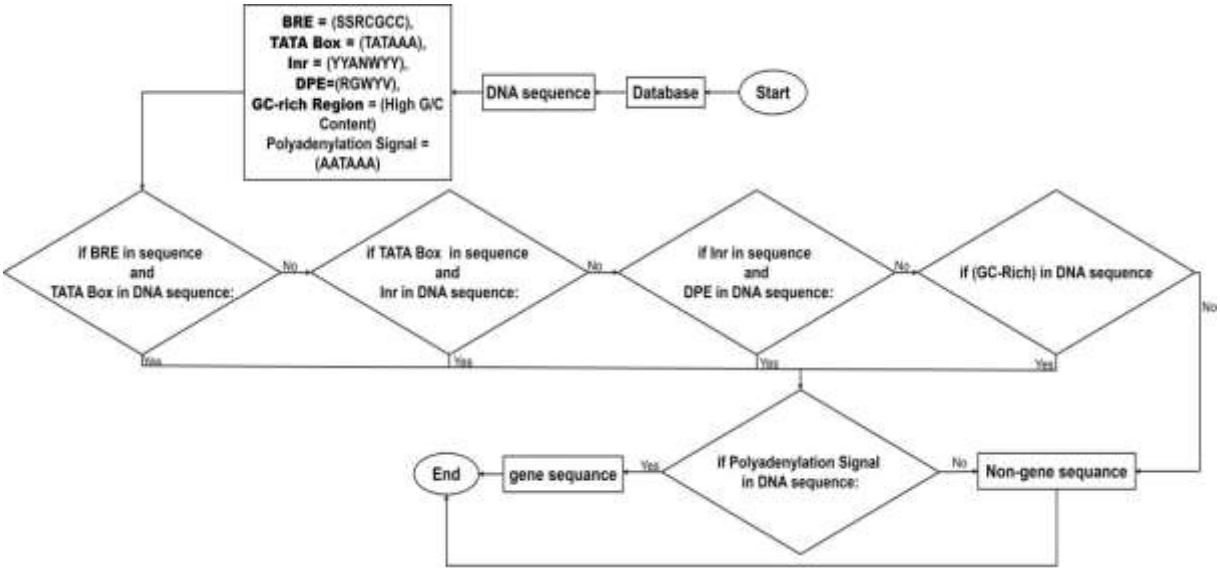

Figure 2: Genomic feature extraction and database storage system.

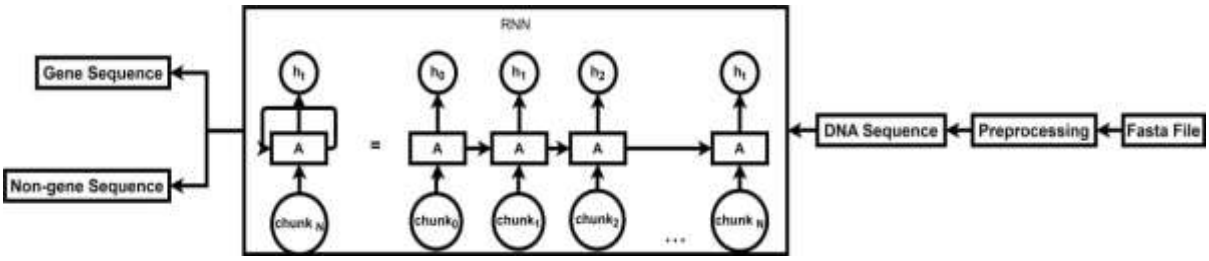

Figure 3: RNN-Core model for Core-promoter classification.

4.2 Core Promoter

Table 2 presents the conservation frequencies of core promoter elements in the provided matrix for the GRCh38.p14 assembly with IDs NC_000001.11, NC_000002.12, NC_000003.12, and NC_000004.12. Each element – BREu, TATA box, BREd, Inr, and DPE – is characterized by high diagonal percentages (98.2%, 96.85%, 98.1%, 98.05%, and 99.5%, respectively), indicating strong sequence specificity. Off-diagonal values, representing minor deviations, are below 1.5%, suggesting minimal functional overlap. These percentages reflect the stringent evolutionary constraints placed on promoter regions to ensure accurate transcription initiation. The TATA box, for instance, is correctly identified in 96.85% of cases, while misclassification as BREu occurs in only 1.16% of instances. Such precision is critical for proper binding of transcription machinery and regulatory fidelity.

Table 3 presents a dataset comprising labeled DNA sequences categorized into "gene-coding (genetic)" and "non-coding (epigenetic)" regions. The sequences, such as "GCCGTCCCGGCTCGCTCACTGCC" and "AGAACTGGGGGTGGACGGAGAGT", enable analysis of functional genomic elements, transcriptional regulation, and structural variations in coding versus regulatory DNA regions.

4.3 Comparison of Model on Grch38.P14

Table 4 presents performance metrics for various models on the GRCh38.p14 genome assembly. The GRU model achieves the highest accuracy (87.04%) and recall (88.98%), while the proposed RNN-Core model demonstrates superior precision (90.53%) and overall accuracy (92.26%). The CNN and LSTM models exhibit comparable performance, whereas the basic RNN underperforms across all metrics. These results highlight the efficacy of advanced architectures in genomic sequence analysis. Figure 4 shows a bar chart comparison of the performance results.

Table 2: Core-promoter elements distributed in the proposed work sequence GRCh38.p14 database.

	BRE^u (%)	TATAbox (%)	BRE^d (%)	Inr (%)	DPE (%)
BRE^u	98.2	1.33	0.1	0.2	0.17
TATAbox	1.16	96.85	0.69	0.1	1.2
BRE^d	0.15	1.08	98.1	0.15	0.52
Inr	0.57	0.15	0.81	98.5	0.42
DPE	0.15	0.2	0.15	0	99.5

Table 3: Example of storage core promoter as two classes (gene, non-gene).

Sequence	Genetic sequences	Epigenetic sequence
GGTCC,…,ACTGC	GGAGA,...,TTGT	AAGGG….,ACAGA
AGACC,…..,CACAT	TATATA,…,GATCA	GATTA,….,CGTATG
CTGGG,….,GGAGA	TCAATA,…,ACAGG	TACCT,..,ATGATT

Table 4: Comparison mode on GRCh38.p14.

Model	Accuracy, (%)	Pression, (%)	Recall, (%)	F1-Score, (%)
CNN [12]	86.69	87.01	86.12	86.56
RNN	68.97	74.72	77.14	75.91
LSTM[12]	85.46	85.70	84.32	85
GRU [13]	87.04	84.92	88.98	86.9
RNN-Core (proposed)	92.26	90.53	87.79	89.13

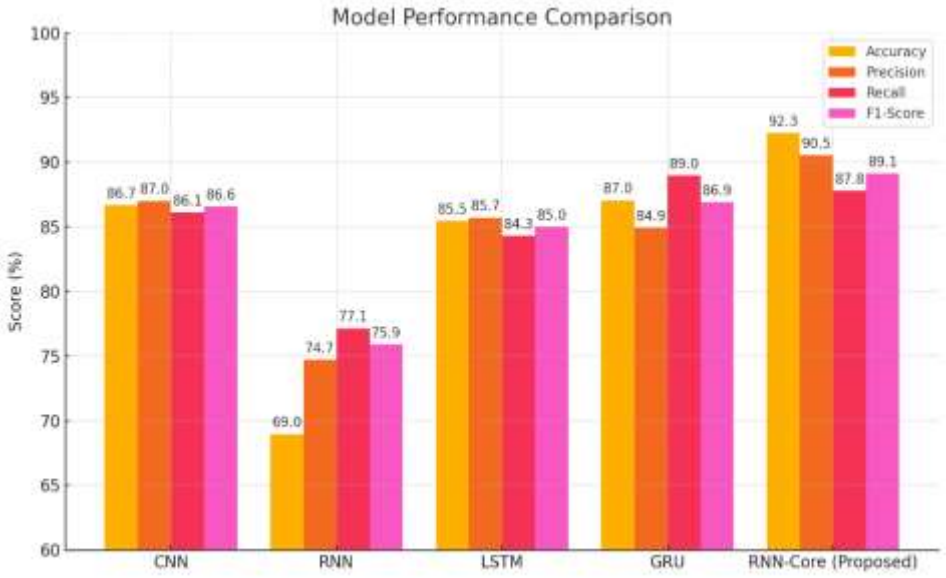

Figure 4: Performance bar of models.

5 CONCLUSIONS

This study successfully developed RNN-based framework that significantly advances genomic sequence interpretation by accurately identifying functional elements through effective modeling of long-range dependencies. By combining the powerful pattern recognition capabilities of deep learning with biological validation, this approach outperforms traditional methods in detecting regulatory motifs and enables standardized, reliable genomic annotation. The framework serves as a bridge between computational techniques and molecular biology, providing practical tools that have important applications in precision medicine and synthetic biology. Given the rapidly increasing volume and complexity of genomic data, such deep learning-based systems will be essential to unlocking the full potential of genomics. Looking ahead, future

enhancements will focus on integrating advanced attention mechanisms and diverse multi-omics datasets to further strengthen the framework's capacity for automated, scalable, and biologically comprehensive analysis. These advancements will open new opportunities for discovery and innovation in the era of big biological data.

REFERENCES

[1] Y. He, Z. Shen, Q. Zhang, S. Wang, and D.-S. Huang, "A survey on deep learning in DNA/RNA motif mining," Briefings in Bioinformatics, vol. 22, no. 4, Jul. 2021, doi: 10.1093/bib/bbaa229.

[2] A. Busia et al., "A deep learning approach to pattern recognition for short DNA sequences," bioRxiv, 2019, doi: 10.1101/353474.

[3] V. X. Jin, G. A. C. Singer, F. J. Agosto-Pérez, S. Liyanarachchi, and R. V. Davuluri, "Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs," BMC Bioinformatics, vol. 7, p. 114, Mar. 2006, doi: 10.1186/1471-2105-7-114.

[4] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–22, 2021, doi: 10.1109/TPAMI.2021.3059968.

[5] J. M. Zhang, "Learning the language of the genome using RNNs," 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:27419365.

[6] T. Shafee and R. Lowe, "Eukaryotic and prokaryotic gene structure," WikiJournal of Medicine, vol. 4, no. 1, pp. 2–6, 2017, doi: 10.15347/wjm/2017.002.

[7] Y. Hara and T. Kawano, "Run-length encoding graphic rules applied to DNA-coded images and animation editable by polymerase chain reactions," Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 19, no. 1, pp. 5–10, 2015, doi: 10.20965/jaciii.2015.p0005.

[8] N. I. Gershenzon and I. P. Ioshikhes, "Synergy of human Pol II core promoter elements revealed by statistical sequence analysis," Bioinformatics, vol. 21, no. 8, pp. 1295–1300, Apr. 2005, doi: 10.1093/bioinformatics/bti172.

[9] S. K. Sønderby, C. K. Sønderby, L. Maaløe, and O. Winther, "Recurrent spatial transformer networks," arXiv preprint arXiv:1506.02025, 2015.

[10] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," SN Computer Science, vol. 2, no. 6, p. 420, 2021, doi: 10.1007/s42979-021-00815-1.

[11] S. Bustin, "Molecular Biology of the Cell, Sixth Edition; ISBN: 9780815344643; and Molecular Biology of the Cell, Sixth Edition, The Problems Book; ISBN 9780815344537," International Journal of Molecular Sciences, vol. 16, no. 12, pp. 28123–28125, Dec. 2015, doi: 10.3390/ijms161226074.

[12] F. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," arXiv preprint arXiv:2305.17473, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258960275.

[13] A. A. K. Nielsen and C. A. Voigt, "Deep learning to predict the lab-of-origin of engineered DNA," Nature Communications, vol. 9, no. 1, p. 3135, 2018, doi: 10.1038/s41467-018-05378-z.

[14] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," IEEE Access, vol. 7, pp. 53040–53065, 2019, doi: 10.1109/ACCESS.2019.2912200.

[15] J. Xiao and Z. Zhou, "Research progress of RNN language model," in Proc. 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2020, pp. 1285–1288, doi: 10.1109/ICAICA50127.2020.9182390.

[16] H. Apaydin et al., "Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting," Water, vol. 12, no. 5, 2020, doi: 10.3390/w12051500.

[17] H. Yin et al., "Rainfall-runoff modeling using long short-term memory based step-sequence framework," Journal of Hydrology, vol. 610, p. 127901, 2022, doi: 10.1016/j.jhydrol.2022.127901.

[18] A. Mathew, P. Amudha, and S. Sivakumari, "Deep learning techniques: An overview," in Advances in Machine Learning Technologies and Applications, A. E. Hassanien, R. Bhatnagar, and A. Darwish, Eds. Singapore: Springer, 2021, pp. 599–608.

[19] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," Journal of Big Data, vol. 8, no. 1, p. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[20] DNA Promoter Region Analyzer. [Online]. Available: https://github.com/jafer0028/DNA-Promoter-Region-Analyzer. Accessed: Jun. 17, 2025.