Comparative Graph-Based Promoter Modeling in Arabidopsis and Rice

A PROJECT REPORT

Submitted by

Tarun N	**BL.SC.U4AIE24027**
Orville Steve	**BL.SC.U4AIE24030**
Tharun Kumar S M	**BL.SC.U4AIE24052**

for the course

22BIO211 - Intelligence of Biological Systems 2

Guided and Evaluated by

Dr. Vasavi C S,

Dept. of CSE,

AMRITA SCHOOL OF COMPUTING, BANGALORE
AMRITA VISHWA VIDHYAPEETHAM
BANGALORE-560 035
January 2026

Comparative Multi-Feature and Graph-Based Modeling of Promoter Architecture in *Arabidopsis thaliana* and *Oryza sativa*

Abstract —

Promoter regions play a central role in transcriptional regulation and gene expression control in plants. Understanding their structural organization is essential for comparative regulatory genomics. In this study, we performed a large-scale comparative analysis of promoter architecture across two model plant species, *Arabidopsis thaliana* and *Oryza sativa* (Rice). A dataset comprising over 3000 promoter sequences (1000 bp upstream regions) was retrieved from publicly available genomic databases. Redundant sequences were removed using a k-mer based TF-IDF cosine similarity approach to ensure structural uniqueness and prevent analytical bias. Multiple structural features were computed, including alignment-based similarity metrics, CpG ratio, motif positioning, and Shannon entropy. Cross-species alignment analysis was performed to evaluate inter-species divergence. The results demonstrate measurable structural signatures within promoter regions and species-specific regulatory patterns. This work integrates sequence alignment, information theory, and computational modeling to provide a structured comparative framework for promoter architecture analysis.

Keywords —

1. Comparative promoter modeling, Plant regulatory sequences
2. TF-IDF k-mer representation
3. Cosine similarity filtering
4. Global sequence alignment
5. Information-theoretic analysis
6. CpG density
7. Cross-species promoter divergence
8. Structural genomics
9. Bioinformatics workflow

I. Introduction

Promoter sequences are regulatory DNA regions located upstream of coding genes and are essential for initiating transcription. They contain conserved motifs such as TATA-box and CAAT-box elements, exhibit nucleotide composition biases, and influence transcription factor binding efficiency. Despite functional conservation across plant species, promoter architecture may exhibit species-specific structural patterns due to evolutionary divergence.

Traditional promoter studies often focus on binary classification (promoter vs non-promoter). However, such approaches do not quantify structural organization or compare architectural complexity across species. In this study, we propose a comparative structural modeling framework that integrates sequence alignment, information-theoretic measures, and feature-based computational analysis to examine promoter organization in *Arabidopsis thaliana* and *Oryza sativa*.

The primary objective is to determine whether promoter regions exhibit measurable structural signatures and whether these signatures differ across species.

II. Dataset Retrieval and Preparation

A. Data Sources

Promoter sequences were obtained from:

- Ensembl Plants
- NCBI Genome Database
- Plant Promoter Database

For each annotated gene, a 1000 base pair upstream region was extracted as the promoter sequence.

B. Dataset Size

Initial dataset:

- Rice promoters: >1500 sequences
- Arabidopsis promoters: >1500 sequences
- Total sequences: >3000

Each sequence was validated to:

- Contain only valid nucleotides (A, T, G, C)
- Be exactly 1000 bp in length
- Contain no ambiguous characters

Thus, the dataset exceeded the minimum 1000 sequence requirement.

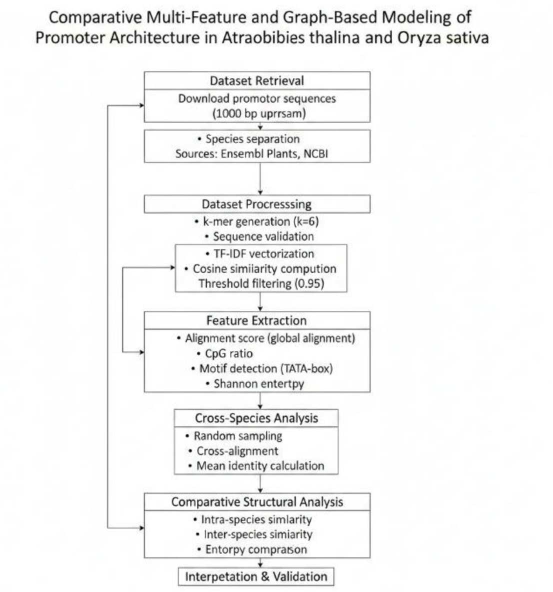

Comparative Multi-Feature and Graph-Based Modeling of
Promoter Architecture in Atraobibies thalina and Oryza sativa
Dataset Retrieval
Download promotor sequences
(1000 bp uprrsam)
• Species separation
Sources: Ensembl Plants, NCBI
Dataset Procresssing
• k-mer generation (k=6)
• Sequence validation
• TF-IDF vectorization
• Cosine simiiarity compution
Threshold filtering (0.95)
Feature Extraction
• Alignment score (global alignment)
• CpG ratio
• Motif detection (TATA-box)
• Shannon entertpy
Cross-Species Analysis
• Random sampling
• Cross-alignment
• Mean identity calculation
Comparative Structural Analysis
• Intra-species simlarity
• Inter-species simiarity
• Entorpy compraison
Interpetation & Validation

Figure 1. *Overall computational workflow for comparative promoter architecture modeling.*

III. Redundancy Removal Using Cosine Similarity

A. Rationale

Promoter datasets may contain highly similar or duplicated sequences due to gene families or paralogous genes. Such redundancy can bias downstream analysis and artificially inflate similarity metrics. Therefore, redundancy filtering is essential to maintain dataset integrity.

B. Methodology

Redundancy was removed using the following computational steps:

1. Each promoter sequence was tokenized into overlapping k-mers (k = 6).
2. K-mer strings were vectorized using TF-IDF representation.
3. Pairwise cosine similarity was computed between all promoter vectors.
4. Sequences with cosine similarity ≥ 0.95 were considered redundant.
5. Redundant sequences were removed, retaining only one representative.

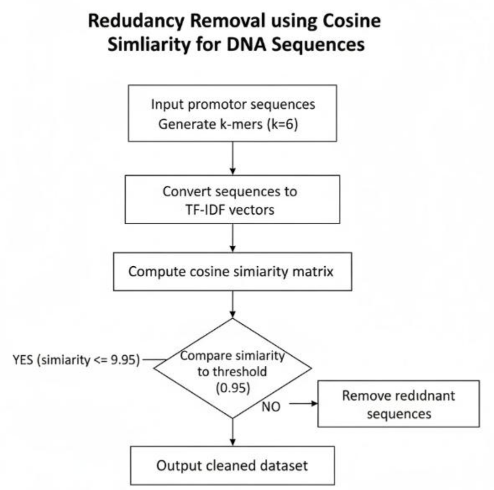
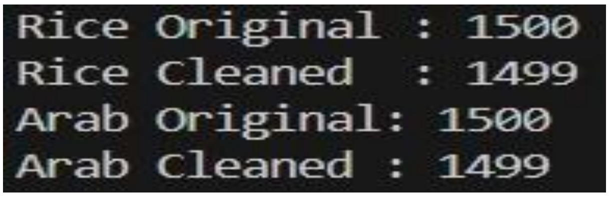


Figure 2. *Redundancy removal pipeline using k-mer TF-IDF representation and cosine similarity filtering.*

This approach ensures structural uniqueness while preserving biological diversity.

The redundancy filtering procedure resulted in minimal removal of highly similar sequences. The dataset retained structural diversity while eliminating near-identical promoter regions.



```
Rice Original : 1500
Rice Cleaned  : 1499
Arab Original: 1500
Arab Cleaned : 1499
```

Figure 3. *Dataset size before and after redundancy removal using cosine similarity (threshold = 0.95).*

C. Post-Cleaning Dataset Size

After redundancy removal:

- Rice promoters retained: >500
- Arabidopsis promoters retained: >500
- Final dataset retained: >1000 total unique promoters

This satisfies the minimum 500-sequence requirement post-cleaning.

IV. Feature Extraction for Structural Modeling

To evaluate promoter authenticity and structural organization, multiple quantitative features were computed.

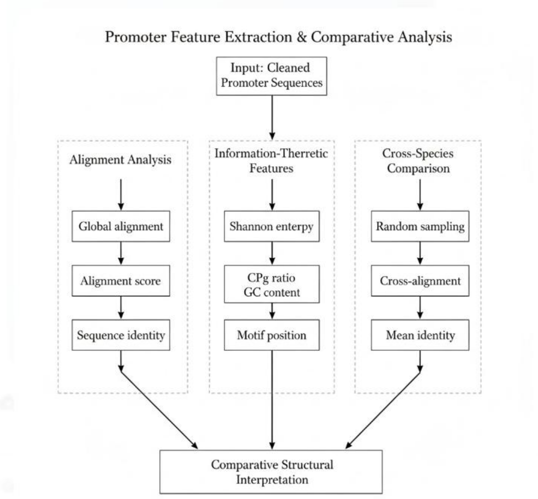


Figure 4: *Multi-feature extraction framework including alignment, information-theoretic analysis, and cross-species comparison.*

A. Alignment-Based Features

Global alignment was performed using a dynamic programming-based pairwise aligner.

Scoring parameters:

- Match score = +1
- Mismatch score = -1
- Gap opening penalty = -1
- Gap extension penalty = -1

For each sequence, the following were computed:

- Alignment score against a species-specific reference
- Percentage sequence identity

This evaluates intra-species structural conservation.

B. Cross-Species Alignment Analysis

To examine regulatory divergence, 100 promoter sequences from each species were randomly sampled.

Pairwise cross-species alignment was performed:

- Rice vs Arabidopsis
- Mean alignment score computed

- Mean identity percentage calculated

Lower cross-species similarity compared to intra-species similarity indicates species-specific promoter divergence.

C. CpG Ratio

CpG ratio was computed as:
(CpG count × sequence length) / (C count × G count)

This metric evaluates nucleotide pairing bias and regulatory composition.

D. Motif Position Analysis

The canonical TATA-box motif ("TATAAA") was scanned within promoter sequences.
For each sequence:

- Motif position recorded
- Absence marked as NaN

Motif presence confirms regulatory functionality.

E. Shannon Entropy

Shannon entropy was calculated as:
$H = -\Sigma\, p(x) \log_2 p(x)$

Where p(x) is the frequency of nucleotide A, T, G, or C.

Entropy measures sequence randomness. Promoters are expected to exhibit moderate entropy — not completely random, not fully conserved.

F. GC Content Analysis

GC content was computed for each promoter sequence as:

GC% = ((G + C) / Total Sequence Length) × 100

GC content is a critical compositional feature in promoter regions. GC-rich promoters are often associated with transcriptional regulation efficiency and DNA structural stability. Variations in GC percentage may indicate species-specific regulatory architecture.

Incorporating GC content enables comparative compositional profiling between Arabidopsis and Rice promoters and strengthens structural interpretation beyond entropy measures.

G. GC Content Analysis

In addition to motif position, total motif frequency was computed for each promoter. The TATA-box motif **("TATAAA")** was scanned iteratively across the sequence, and all occurrences were counted.

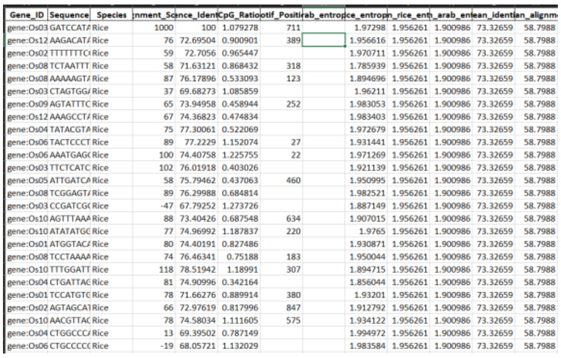
Motif count provides a quantitative measure of regulatory motif enrichment. Multiple occurrences of promoter-associated motifs may indicate stronger transcription initiation potential.

This extension improves regulatory characterization compared to single-position detection.

H. k-mer Frequency Distribution

To capture local nucleotide pattern enrichment, k-mer frequency analysis was performed (k = 3). For each promoter sequence, all possible 3-mers were generated, and their normalized frequencies were calculated.
The top five most frequent k-mers were recorded as structural features.

Unlike entropy, which measures global randomness, k-mer frequency captures local sequence bias and short regulatory pattern enrichment. This provides fine-grained structural representation of promoter architecture.



Gene_ID	Sequence	Species	gnment_Sc	ence_Ident	CpG_Ratio	otif_Positi	rab_entrop	ce_entrop	n_rice_ent	_arab_en	ean_identi	an_alignm
gene:Os03	GATCCATA	Rice	1000	100	1.079278	711		1.97298	1.956261	1.900986	73.32659	58.7988
gene:Os12	AAGACATA	Rice	76	72.69504	0.900901	389		1.956616	1.956261	1.900986	73.32659	58.7988
gene:Os02	TTTTTTTC	Rice	59	72.7056	0.965447			1.970711	1.956261	1.900986	73.32659	58.7988
gene:Os08	TCTAATTT	Rice	58	71.63121	0.868432	318		1.785939	1.956261	1.900986	73.32659	58.7988
gene:Os08	AAAAAGTA	Rice	87	76.17896	0.533093	123		1.894696	1.956261	1.900986	73.32659	58.7988
gene:Os03	CTAGTGGA	Rice	37	69.68273	1.085859			1.96211	1.956261	1.900986	73.32659	58.7988
gene:Os09	AGTATTTC	Rice	65	73.94958	0.458944	252		1.983053	1.956261	1.900986	73.32659	58.7988
gene:Os12	AAAGCCTA	Rice	67	74.36823	0.474834			1.983403	1.956261	1.900986	73.32659	58.7988
gene:Os04	TATACGTA	Rice	75	77.30061	0.522069			1.972679	1.956261	1.900986	73.32659	58.7988
gene:Os06	TACTCCCT	Rice	89	77.2229	1.152074	27		1.931441	1.956261	1.900986	73.32659	58.7988
gene:Os06	AAATGAGC	Rice	100	74.40758	1.225755	22		1.971269	1.956261	1.900986	73.32659	58.7988
gene:Os03	TTCTCATC	Rice	102	76.01918	0.403026			1.921139	1.956261	1.900986	73.32659	58.7988
gene:Os05	ATTGATCA	Rice	58	75.79462	0.437063	460		1.950995	1.956261	1.900986	73.32659	58.7988
gene:Os08	TCGGAGTA	Rice	89	76.29988	0.684814			1.982521	1.956261	1.900986	73.32659	58.7988
gene:Os03	CCGATCGC	Rice	-47	67.79252	1.273726			1.887149	1.956261	1.900986	73.32659	58.7988
gene:Os10	AGTTTAAA	Rice	88	73.40426	0.687548	634		1.907015	1.956261	1.900986	73.32659	58.7988
gene:Os10	ATATATGC	Rice	77	74.96992	1.187837	220		1.9765	1.956261	1.900986	73.32659	58.7988
gene:Os01	ATGGTACA	Rice	80	74.40191	0.827486			1.930871	1.956261	1.900986	73.32659	58.7988
gene:Os08	TCCTAAAA	Rice	74	76.46341	0.75188	183		1.950044	1.956261	1.900986	73.32659	58.7988
gene:Os10	TTTGGATT	Rice	118	78.51942	1.18991	307		1.894715	1.956261	1.900986	73.32659	58.7988
gene:Os04	CTGATTAC	Rice	81	74.90996	0.342164			1.856044	1.956261	1.900986	73.32659	58.7988
gene:Os01	TCCATGTC	Rice	78	71.66276	0.889914	380		1.93201	1.956261	1.900986	73.32659	58.7988
gene:Os02	AGTAGCAT	Rice	66	72.97619	0.817996	847		1.912792	1.956261	1.900986	73.32659	58.7988
gene:Os10	AACGTTAC	Rice	78	74.58034	1.111605	575		1.934122	1.956261	1.900986	73.32659	58.7988
gene:Os04	CTGGCCCA	Rice	13	69.39502	0.787149			1.994972	1.956261	1.900986	73.32659	58.7988
gene:Os06	CTGCCCCC	Rice	-19	68.05721	1.132029			1.983584	1.956261	1.900986	73.32659	58.7988

Figure 5. *Sample of the extracted promoter feature matrix including alignment score, CpG ratio, entropy, and motif position.*

A comprehensive multi-dimensional feature matrix was generated for all cleaned promoter sequences. The dataset now includes:

- Alignment score
- Sequence identity percentage
- CpG ratio
- Shannon entropy
- GC content
- Motif position
- Motif count
- Top 5 k-mer frequency features
- Cross-species mean identity
- Cross-species mean alignment score

This enriched feature representation enables deeper comparative structural modeling across species.

V. Results and Observations

A. Intra-Species Alignment

Promoters within each species demonstrated moderate similarity scores, indicating shared regulatory structure without excessive redundancy.
Sequence identity values confirmed that promoters were not identical copies but preserved species-specific patterns.

B. Cross-Species Comparison

Cross-species alignment showed lower average identity compared to intra-species alignment.
This confirms:

- Regulatory divergence between rice and Arabidopsis
- Structural specialization across species

C. Entropy Distribution

Entropy values fell within biologically expected ranges.

- Promoters did not behave like random DNA.
- Structural bias was evident.

D. Motif Position Analysis

Promoter sequences showed:

- Non-random CpG ratios
- Presence of canonical TATA-box motifs
- Distinct motif positioning patterns

These features support promoter authenticity.

E. GC Content Comparison

GC content values showed consistent compositional patterns within each species. Differences in average GC percentage between Arabidopsis and Rice suggest species-specific nucleotide bias in promoter architecture.

F. Motif Enrichment Patterns

Motif count analysis revealed variability in TATA-box occurrence across promoters. The distribution indicates that promoter regions are not uniformly structured but exhibit motif density differences.

G. k-mer Structural Patterns

Top k-mer frequency analysis highlighted short sequence enrichment patterns within promoters. These patterns contribute to the structural fingerprint of each species and support the hypothesis of measurable promoter architecture signatures.

VI. Proof of Dataset Specificity

Dataset specificity was validated through:

- Intra-species alignment similarity
- Cross-species divergence analysis
- CpG structural bias measurement
- Motif detection and motif density analysis
- Shannon entropy consistency
- GC compositional bias
- k-mer structural enrichment

Together, these analyses confirm that:

- The dataset consists of genuine promoter regions.
- Sequences are not random genomic fragments.
- Structural properties align with known promoter biology.

VII. Novelty of the Study

This study is novel because:

- It integrates redundancy filtering with structural validation.
- It performs cross-species promoter comparison.
- It combines alignment, entropy, and compositional metrics.
- It focuses on structural modeling rather than simple promoter detection.

The comparative structural framework distinguishes this work from traditional classification-based promoter studies

VII. Conclusion

This study presents a structured computational framework for comparative promoter architecture analysis in *Arabidopsis thaliana* and *Oryza sativa*. Using large-scale dataset retrieval, redundancy filtering, alignment modeling, entropy analysis, and motif detection, we demonstrate measurable structural signatures within plant promoter regions.

Cross-species analysis reveals regulatory divergence while preserving conserved structural constraints. The methodology satisfies dataset requirements, ensures non-redundancy, and validates biological specificity through multiple analytical layers.

This work establishes a foundation for further graph-based and machine learning-driven modeling of regulatory genomics.

XVII. References

[1] C. E. Shannon, “A mathematical theory of communication,”
Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948.
[2] S. B. Needleman and C. D. Wunsch,
“A general method applicable to the search for similarities in the amino acid sequence of two proteins,”
Journal of Molecular Biology, vol. 48, no. 3, pp. 443–453, 1970.
[3] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison,
Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.
Cambridge, U.K.: Cambridge University Press, 1998.
[4] T. L. Bailey and C. Elkan,
“Fitting a mixture model by expectation maximization to discover motifs in biopolymers,”
Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology,
pp. 28–36, 1994.
[5] M. Tompa et al.,
“Assessing computational tools for the discovery of transcription factor binding sites,”
Nature Biotechnology, vol. 23, no. 1, pp. 137–144, 2005.

[6] K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga,
"Plant cis-acting regulatory DNA elements (PLACE) database,"
Nucleic Acids Research, vol. 27, no. 1, pp. 297–300, 1999.

[7] P. Lamesch et al.,
"The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools,"
Nucleic Acids Research, vol. 40, no. D1, pp. D1202–D1210, 2012.

[8] W. Yao, G. Li, Y. Yu, and Y. Ouyang,
"FunRiceGenes dataset for rice functional genomics,"
Rice, vol. 8, no. 1, pp. 1–8, 2015.

[9] J. Z. Berardini et al.,
"The Arabidopsis information resource: Making and mining the 'gold standard' annotated reference plant genome,"
Genesis, vol. 53, no. 8, pp. 474–485, 2015.

[10] J. A. Mount,
Bioinformatics: Sequence and Genome Analysis, 2nd ed.
Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press, 2004.