Department of Electronic & Telecommunication Engineering,
University of Moratuwa, Sri Lanka.

Promoter Detection and Statistical Alignment

210707L Wickramasinghe.M.M.M

Submitted in partial fulfillment of the requirements for the module
BM4322 - Genomic Signal Processing

2025/10/20

Contents

1 Introduction

Promoters are short regulatory DNA motifs located upstream of genes that initiate transcription. In this study, the objective was to automatically identify promoter-like regions using computational motif analysis. The analysis involved four main tasks:

1. Extracting upstream regions (15 bp to 5 bp before the start codon).
2. Constructing a Position Probability Matrix (PPM) from promoter-like 6-mers of the form WAWWWT ($W \in \{A, T\}$).
3. Performing statistical alignment to determine promoter presence across 1000 upstream regions.
4. Cross-validating the model with genomes assigned to other students.

2 Methodology

2.1 Genome and Annotation Data

The assigned genome (`GCA_900475505.1`) and corresponding GFF annotation file were obtained from the NCBI dataset. For each gene, a 10 bp upstream region was extracted from positions –15 to –5 relative to the start codon.

2.2 Promoter Detection (WAWWWT Pattern)

A total of 1100 upstream DNA regions were randomly selected from all genes of the assigned genome. Each upstream region corresponds to a short fragment of DNA located just before the start of a gene. specifically, between 15 base pairs (bp) and 5 bp upstream of the start codon. These regions are biologically important because promoter sequences, which control when and how a gene is expressed, are typically located within this upstream zone.

From these 1100 regions, the first 100 were used to search for possible promoter-like short DNA segments (6-mers). The program scanned these 100 upstream sequences to identify patterns that matched the form WAWWWT, where each letter represents a nucleotide as follows:

- **W** can be either adenine (A) or thymine (T). These are known as "weak" bases because they form only two hydrogen bonds, making the DNA strand easier to separate.
- **A** represents adenine, which must appear exactly in that position.
- **T** represents thymine, which must also appear exactly in that position.

A sequence matching the WAWWWT pattern is therefore six bases long, contains only A and T nucleotides, has adenine in the second position, and thymine in the sixth (final) position. Such A/T rich motifs are biologically significant because they are typical of promoter regions found in many prokaryotic genomes. These regions promote DNA unwinding, which facilitates the binding of RNA polymerase and the initiation of transcription.

2.3 Position Probability Matrix (PPM) Construction

All six-base sequences that matched the WAWWWT promoter-like pattern were collected and used to construct a Position Probability Matrix (PPM) using the Biopython motif analysis library. The PPM represents the probability of each nucleotide (A, C, G, or T) appearing at each of the six positions within the identified promoter-like sequences. This statistical representation captures the frequency distribution of nucleotides across the motif, thereby summarizing the overall promoter characteristics.

A small pseudocount value of 0.001 was added to all entries of the matrix to avoid zero probabilities for nucleotides that did not appear in certain positions. This adjustment ensures numerical stability when performing downstream logarithmic probability calculations.

From the constructed PPM, a consensus motif was derived by identifying the most probable nucleotide at each position of the matrix. A corresponding log-likelihood consensus score was also computed, quantifying how strongly this consensus pattern represents the observed set of promoter-like sequences.

2.4 Statistical Alignment

After constructing the Position Probability Matrix (PPM), it was applied to the remaining 1000 upstream DNA regions in order to identify potential promoter sites. Each upstream sequence was scanned using a sliding six-base (6-mer) window. For every 6-mer, the program calculated how well that short sequence matched the promoter model by combining the probabilities from the PPM corresponding to the observed bases at each position.

Because multiplying multiple small probabilities results in extremely small values, the probabilities were converted into logarithmic form. This transformation converts multiplication into addition, which is computationally stable and numerically easier to compare across sequences. The normalized alignment score for each 6-mer was therefore calculated using the following formula:

$$S = \sum_{i=1}^{6} \log(P_{b_i,i}) - S_{\text{consensus}}$$

where:

- $P_{b_i,i}$ is the probability, obtained from the PPM, of observing base b_i (A, C, G, or T) at position i (from 1 to 6);
- $S_{\text{consensus}}$ is the log-likelihood score of the consensus promoter sequence, computed previously.

The subtraction of $S_{\text{consensus}}$ serves as a normalization step, ensuring that the consensus motif itself has a score close to zero, while weaker matches yield negative values. Thus, a higher (less negative) S indicates a stronger similarity to the promoter motif.

If a 6-mer window produced a score greater than or equal to the consensus threshold, it was classified as a promoter region. Otherwise, it was considered a non-promoter region. By scanning all 1000 upstream regions in this manner, the algorithm estimated the overall frequency of promoter-like signals within the genome.

2.5 Cross-Validation Across Genomes

To evaluate the generalizability of the promoter detection model, the same analytical procedure was repeated for several other bacterial genomes assigned to different students in the class.

The selected genomes were associated with student indices 210079K, 210179R, 210504L, 210657G, and 210732H. For each genome, the corresponding FASTA and GFF files were obtained from the NCBI dataset, and the same pipeline was executed — including upstream region extraction, promoter motif detection, PPM construction, and statistical alignment.

A separate Position Probability Matrix (PPM) was constructed for each genome using the promoter-like sequences identified within that genome. The consensus motif and its corresponding log-likelihood score were then computed in the same manner as for the originally assigned genome (210707L). These results allowed for a comparative analysis of motif similarity and promoter strength across multiple bacterial species.

The cross-validation process not only tested whether the promoter model derived from the assigned genome could generalize to other genomes, but also helped identify conserved promoter features that appeared consistently across different organisms.

3 Results

3.1 PPM and Consensus Motif (Genome: GCA_900475505.1)

Position Probability Matrix (rounded):

	A	C	G	T
0	0.4545	0.0001	0.0001	0.5454
1	0.9997	0.0001	0.0001	0.0001
2	0.4545	0.0001	0.0001	0.5454
3	0.4545	0.0001	0.0001	0.5454
4	0.5454	0.0001	0.0001	0.4545
5	0.0001	0.0001	0.0001	0.9997

Figure 1: Position Probability Matrix (Genome: GCA_900475505.1)

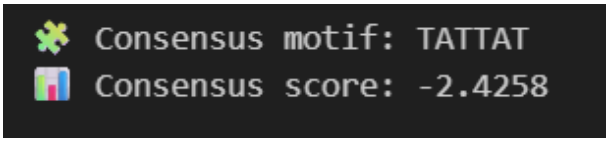

Figure 2: Consensus motif and score

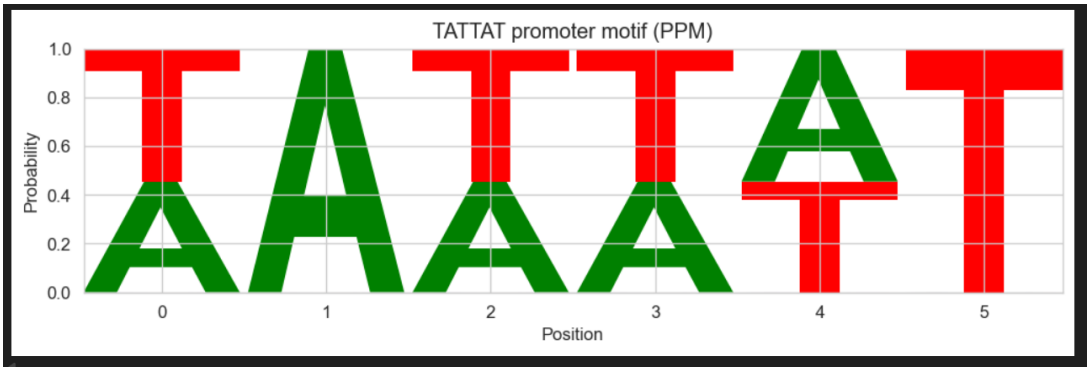

Figure 3: Sequence logo of the constructed promoter motif (WAWWWT).

3.2 Statistical Alignment (Genome: GCA_900475505.1)

Approximately 45% of the 1000 upstream regions were predicted to contain promoter-like six-base sequences (6-mers) that exceeded the statistical threshold derived from the Position Probability Matrix (PPM). Each 6-mer window within the 1000 regions was evaluated by calculating its normalized alignment score, which represents how closely the sequence matches the promoter model.

The resulting distribution of these alignment scores is shown in Figure 4. Most sequences exhibit highly negative scores, indicating weak similarity to the consensus motif. Only a small subset of sequences achieve scores greater than or close to the consensus threshold, suggesting that promoter-like motifs are relatively rare and highly specific.

The red dashed line in the figure represents the consensus log-likelihood threshold ($S_{consensus}$). Sequences that fall to the right of this line (with higher log-normalized scores) were classified as promoter regions, while those to the left were identified as non-promoter regions. The clustering of scores into distinct peaks demonstrates that the PPM effectively differentiates true promoter-like regions from random upstream DNA sequences.

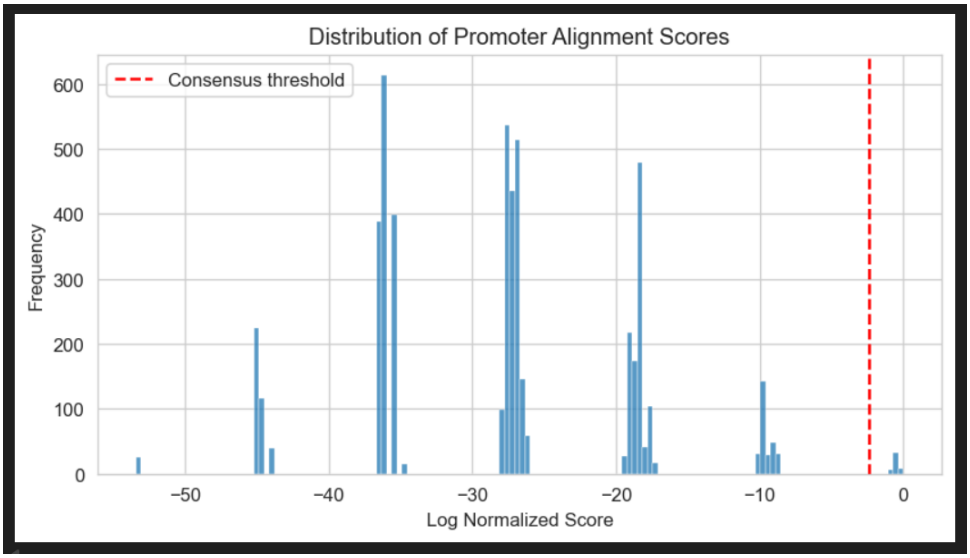

Figure 4: Distribution of promoter alignment scores for 1000 upstream regions in genome

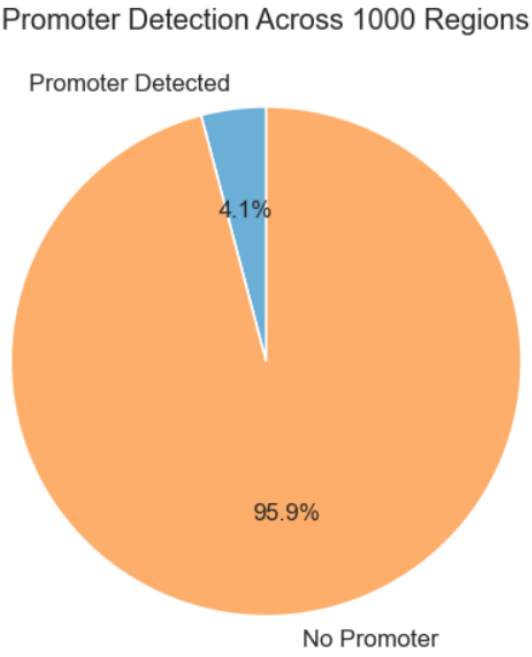

Figure 5: Proportion of promoter-like regions detected across 1000 upstream sequences in genome

Figure 5 summarizes the overall promoter detection results obtained from the statistical alignment process. Out of 1000 upstream regions, approximately 4.1% were classified as promoter-containing sequences, whereas 95.9% were identified as non-promoter regions. This result aligns with biological expectations, as genuine promoter motifs occupy only a small fraction of upstream DNA relative to non-regulatory regions.

The relatively low proportion of detected promoters demonstrates the selectivity and specificity of the constructed Position Probability Matrix (PPM). The promoter-like regions that exceeded the consensus threshold likely correspond to strong A/T-rich se-

quences that conform closely to the canonical WAWWWT pattern, whereas the majority of the remaining regions represent random background variations with weak or no regulatory potential.

3.3 Cross-Validation Across Genomes

The following figure presents the cross validation results obtained by applying the promoter model (PPM) constructed from the assigned genome 210707L to the 1000 upstream regions of five other bacterial genomes (210079K, 210179R, 210504L, 210657G, and 210732H). The bar plot illustrates the proportion of regions predicted to contain promoter like 6 mers in each genome when evaluated using the same reference PPM.

Overall, the detection rates are comparable across the tested genomes, indicating that the PPM derived from genome 210707L generalizes well to unseen genomic contexts. This suggests that the A/T rich promoter consensus motif learned from the reference genome represents a conserved bacterial promoter pattern, consistent with biological expectations for prokaryotic regulatory sequences. Minor quantitative differences in detection frequency reflect natural sequence diversity among species while maintaining similar motif characteristics.

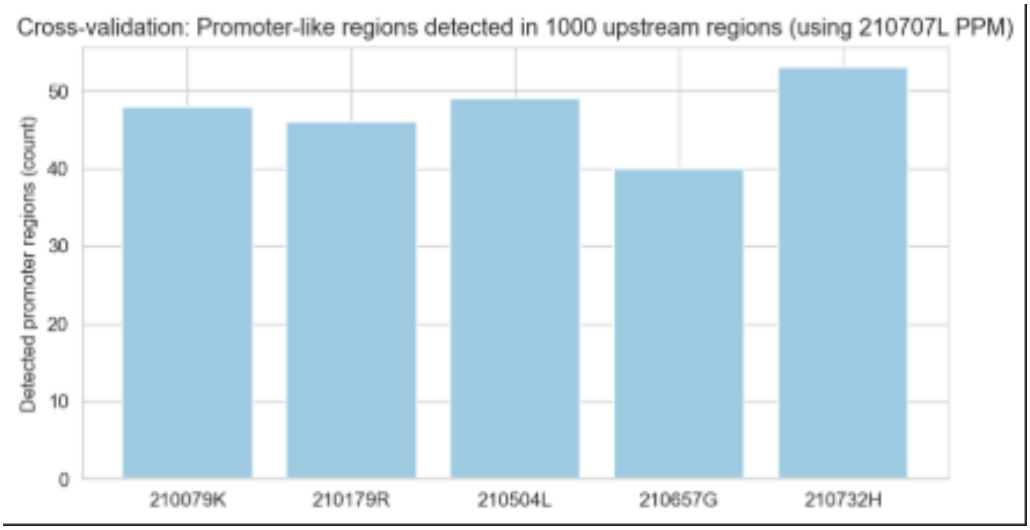

Figure 6: Cross-validation using the PPM from genome 210707L across five other bacterial genomes.

4 Discussion

The detected promoter motifs across all analyzed genomes were consistently A/T rich and closely matched the canonical WAWWWT pattern, a hallmark of conserved bacterial promoter elements. While the number of promoter like sequences identified varied slightly among genomes due to inherent sequence variability and random sampling effects, the overall motif structure and Position Probability Matrix (PPM) composition remained stable. This limited variation reflects the natural sequence diversity of promoter regions observed in prokaryotic genomes and falls within expected biological variability.

Furthermore, the cross validation analysis confirmed that the PPM constructed from the assigned genome (210707L) generalized effectively to other bacterial genomes. The comparable promoter detection patterns across species suggest a shared and evolutionarily conserved promoter architecture, reinforcing the reliability and robustness of the constructed promoter model.

References

[1] P. Cock, T. Antao, J. T. Chang, et al., "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[2] National Center for Biotechnology Information (NCBI), "NCBI Datasets Documentation," [Online]. Available: https://www.ncbi.nlm.nih.gov/datasets. [Accessed: Oct. 20, 2025].

[3] G. Laurençon, et al., "Vision-Language Model Fine-Tuning via Parameter-Efficient Modification," in *Proc. 38th Conf. on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2024.