UNIVERSITY OF MORATUWA, SRI LANKA

Faculty of Engineering

Department of Electronic and Telecommunication Engineering

Semester 7 (Intake 2020)

BM4321 Genomic Signal Processing

Promoter Discovery in Bacteria

Azotobacter vinelandii

A. T. P. Amarasekara

200023C

Table of Contents

List of Figures

Introduction

Promoters are essential DNA sequences that regulate gene transcription by providing binding sites for RNA polymerase. Identifying promoters in bacterial genomes, particularly in nitrogen-fixing bacteria of the genus *Azotobacter*, helps understand gene regulation mechanisms and optimize their applications in agriculture and biotechnology.

This report focuses on alignment of promoters in bacterial genomes using different aligning algorithms. It includes the detection of promoters using the local alignment with WWWW motif, analyzing their characteristics, and performing statistical alignment using the position probability matrix. Specifically, the first genome (*Azotobacter vinelandii*, accession CP001157.1) was selected for analysis.

Methodology

Data Acquisition

Genome assembly data for Azotobacter vinelandii (accession CP001157.1) was downloaded from the NCBI GenBank database. Two files were obtained:

- The .fna file containing the genome sequence.
- The .gtf file providing the genomic loci of genes.

```
ID: CP001157.1
Name: CP001157.1
Description: CP001157.1 Azotobacter vinelandii DJ, complete genome
Number of features: 0
Seq('TTTATAGGGAAGCTTGCCGATCCATGTGGATAACCCTGGTCCGGACGGATACAA...GGC')
```

Figure 1- Genome details of Azotobacter vinelandii, *accession CP001157.1*

The 100 bases upstream and 3 bases downstream of coding sequences on the sense strand were extracted for each gene. A total of 2500 sequences were extracted corresponding to the available genes and this number could be verified using the number of genes available in the .gtf file.

```
Number of sequences: 2500
Length of each sequence: [1540 1207 1201 ...  967  184  223]
```

Figure 2- Details of extracted sequences in the sense strand

These sequences were used for further analysis using the local alignment and statistical alignment algorithms.

Verification of Sequences

To ensure accuracy, the extracted sequences were verified by checking for the presence of the start codon Methionine (ATG) at the beginning of the coding region (base pairs from 101st to 103rd location).

```
Gene Avin_00010 does NOT start with Methionine.
Gene Avin_00020 starts with Methionine!
Gene Avin_00030 starts with Methionine!
Gene Avin_00040 starts with Methionine!
Gene Avin_00050 starts with Methionine!
Gene Avin_00120 starts with Methionine!
Gene Avin_00130 starts with Methionine!
Gene Avin_00180 starts with Methionine!
Gene Avin_00190 starts with Methionine!
Gene Avin_00200 starts with Methionine!
Gene Avin_00220 starts with Methionine!
Gene Avin_00230 starts with Methionine!
Gene Avin_00250 starts with Methionine!
Gene Avin_00270 starts with Methionine!
Gene Avin_00290 starts with Methionine!
Gene Avin_00320 does NOT start with Methionine.
Gene Avin_00330 starts with Methionine!
Gene Avin_00360 starts with Methionine!
Gene Avin_00390 starts with Methionine!
Gene Avin_00410 starts with Methionine!
Gene Avin_00420 starts with Methionine!
Gene Avin_00450 starts with Methionine!
Gene Avin_00460 starts with Methionine!
Gene Avin_00470 starts with Methionine!
Gene Avin_00480 starts with Methionine!
...
Gene Avin_52320 starts with Methionine!
Gene Avin_52380 starts with Methionine!
Gene Avin_52490 does NOT start with Methionine.
Number of genes that start with Methionine: 2061
```

Figure 3- Detection of Methionine codon

This resulted in detection of Methionine in 2061 sequences out of the 2500 extracted sequences.

1. Local Alignment of WWWW promoter

Local alignment algorithm was employed to identify the "WWWW" promoter in each of the 2500 genomic sequences. The first 88 base pairs of the upstream sequence of each gene were selected to ensure that the alignment covered a sufficient region, given that multiple local alignments might occur. A gap of 12 base pairs upstream of the gene was chosen to allow sufficient codons (assumed to be 4) for transcription.

To facilitate the promoter search, all "T" and "A" bases within the selected region were replaced with "W" bases. The algorithm applied a match score of 1 and a gap penalty of -2, ensuring that intact nature of the promoter is prioritized.

The local alignment algorithm was implemented to search for the promoter sequence, considering the last position of the maximum score as the potential alignment. After performing the alignment across all 2500 sequences, only those sequences with the exact "WWWW" alignment were considered as having detected a promoter. As a result, **1678 sequences** were identified as containing the promoter, resulting in a promoter detection percentage of **67.12%**. The following histogram presents the distribution of the detected promoter positions across the sequences.

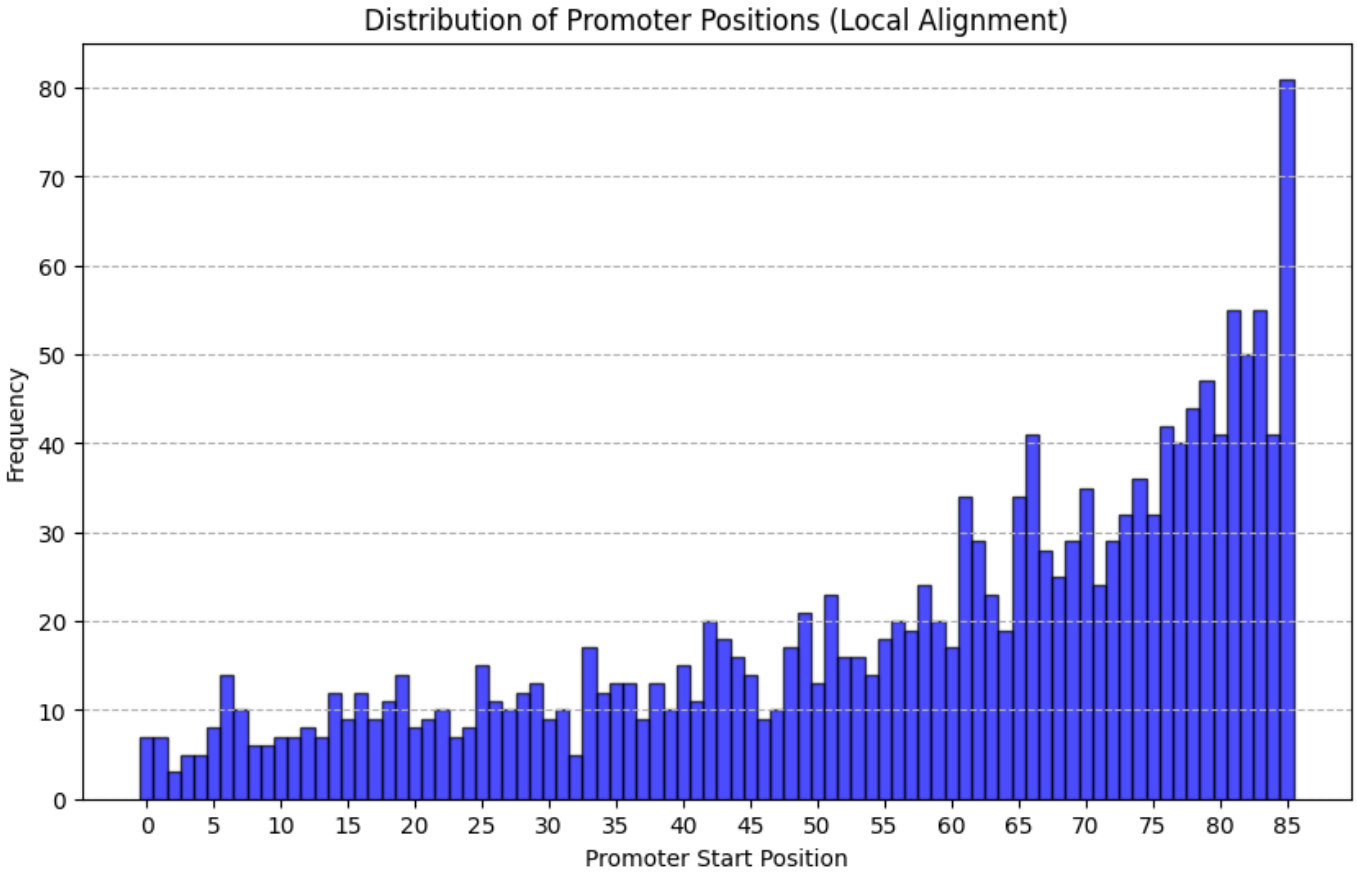

Figure 4- Distribution of the Promoter position in Local alignment

According to the histogram, it is evident that the majority of the detected promoters align in the latter regions of the sequences. This distribution is largely influenced by the method used in the local alignment process, where the last alignment obtained, based on the maximum score, was considered the correct alignment.

2. Extension of the promoter with consecutive “W”s

After detecting the promoter positions, each sequence was further analyzed for consecutive “W”s both forward and backward from the detected promoter location. The total number of consecutive “W”s was recorded for each detected promoter, which was considered as the promoter length. This assessed the length of each promoter by counting the uninterrupted sequence of “W” bases in both directions from the identified promoter region.

A histogram was then plotted to visualize the distribution of the promoter lengths across all the detected promoters. This histogram shows the variation in promoter lengths, revealing patterns and trends in the sequences where the "WWWW" promoter motif is found.

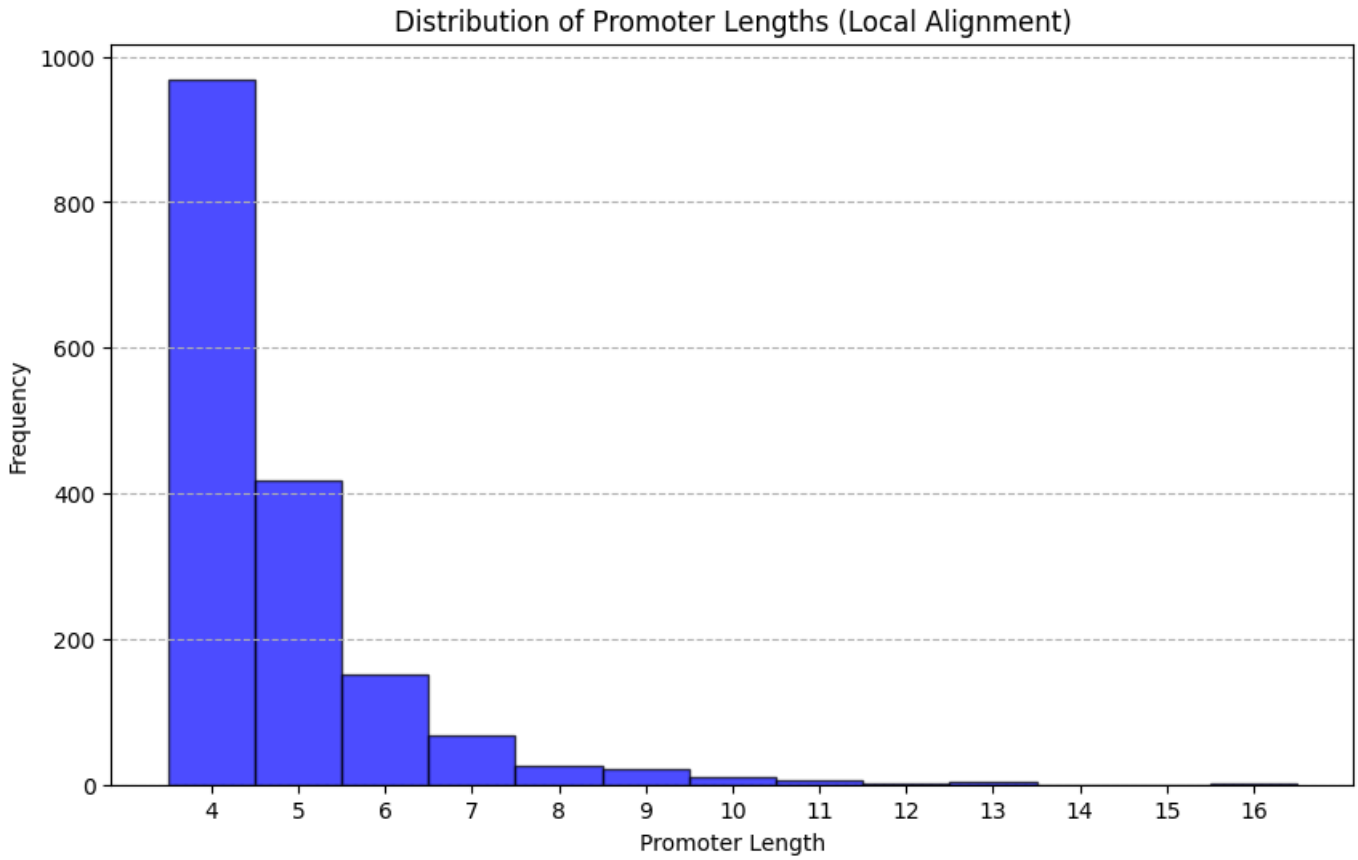

Figure 5- Distribution of Promoter lengths with consecutive Ws

3. Position Probability Matrix

A position probability matrix (PPM) was calculated by considering 6-base long sequences at each of the detected promoter positions in the 1648 sequences. The PPM was constructed by evaluating the frequency of each nucleotide (A, T, C, G) at each of the 6 positions relative to the detected promoter location.

```
['TTTTGC' 'TATTCG' 'TAAACC' ... 'ATAACC' 'AATTGA' 'ATTTCC']
```

Figure 6- Promoter sequences used for PPM

This step was performed without considering the C/G content, meaning that the calculation focused solely on the nucleotide distributions at each position within the 6-base window. The resulting matrix reflects the probability of each nucleotide occurring at each of the 6 positions across all detected promoters.

```
Position Probability Matrix (PPM):
      Pos 1     Pos 2     Pos 3     Pos 4     Pos 5     Pos 6
A  0.564297  0.488913  0.478822  0.383257  0.009306  0.218837
C  0.000996  0.000996  0.000996  0.000996  0.616531  0.359514
G  0.000996  0.000996  0.000996  0.000996  0.367824  0.269885
T  0.433711  0.509095  0.519185  0.614751  0.006338  0.151764
```

Figure 7- PPM for 6 positions

As the exact alignment of the "W" motif is known in the first 4 positions for all the promoter sequences considered for the Position Probability Matrix (PPM), the probabilities for A (adenine) and T (thymine) are higher in these positions. In the 5th position, however, the nucleotide C (cytosine) has the highest probability, followed by G (guanine). In the final (6th) position, the probabilities for all four bases (A, T, C, G) are distributed almost uniformly, indicating a lack of strong preference for any particular nucleotide at this position. This could suggest that the 6th position is more variable and does not have a clear importance for the promoter.

Then, the consensus sequence and the consensus score corresponding to the PPM were obtained.

```
Consensus Sequence: ATTTCC
Consensus Score: -3.8959753941235276
```

Figure 8- Consensus sequence and score

4. Statistical Alignment using the PPM

The statistical alignment algorithm performs a sequence comparison to identify the best matching promoter region based on a consensus sequence. A 6 base window is extracted from the sequence and its norm score (alignment score using the PPM - consensus score) is then calculated. This is done for all possible positions in the sequence. Once all alignments are evaluated, if the best score is below a predefined threshold, the alignment is discarded. Otherwise, the best alignment, its position, and the associated score is obtained.

The number of sequences with detected promoters is influenced by the predefined threshold used in the statistical alignment algorithm. To explore this relationship, the threshold was varied from -1 to -7. When the threshold was set between -3 and -6, the percentage of detected promoters remained nearly constant, providing consistent results. However, when the threshold was set to -7, there was a significant increase in the detection rate, rising from 66.28% to 96.56%. This sharp increase suggests a threshold for effectively distinguishing between potential promoter sequences and other random 6-base pair permutations, as it clearly separates the norm score values associated with promoters from non-promoter sequences.

For further analysis a threshold of -6 was used and the percentage of promoter detection of **66.28%** and the distribution of the promoter positions were obtained.

```
Aligned Promoters: ['ATAACC' 'TATTCG' 'TAAACC' ... 'ATAACC' 'AATTGA' 'ATTTCC']
Alignment Positions: [29 26 79 ... 70 82 71]
Total number of genes with alignments: 1657
Percentage of genes with alignments: 66.28%
```

Figure 9- Details of the results of the Statistical alignment of Promoters for a norm score threshold of -6

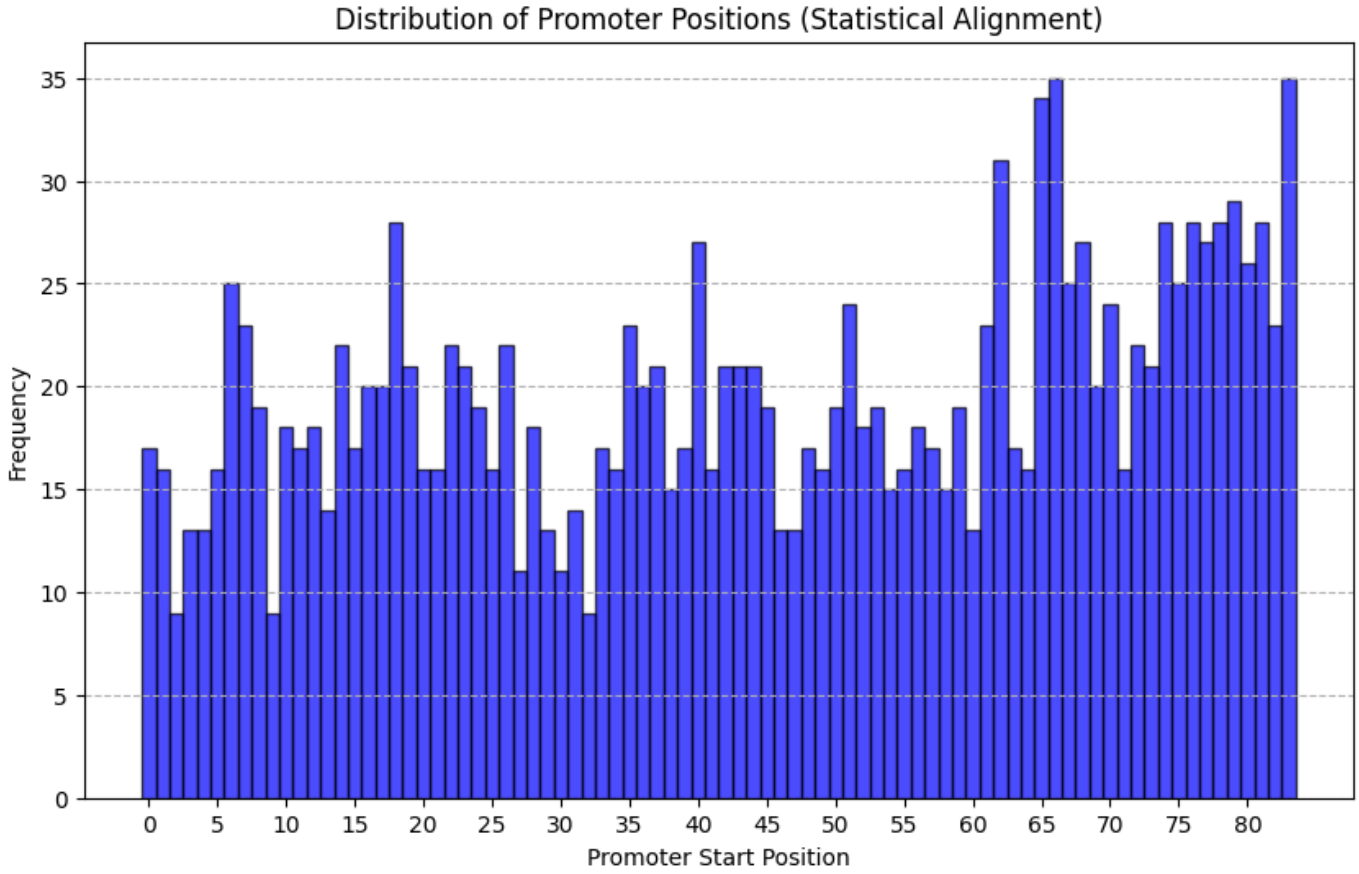

Figure 10- Distribution of the Promoter position in Local alignment (norm score threshold of -6)

5. Comparison of Local Alignment and Statistical Alignment

The local alignment algorithm detected promoters in **67.12%** of sequences, identifying **1678** promoters, while the statistical alignment algorithm detected promoters in **66.28%** of sequences, identifying **1657** promoters. These values are nearly identical, demonstrating that both algorithms are capable of detecting the presence of the "WWWW" promoter.

However, when comparing the distributions of detected promoter positions, clear differences emerge between the two methods. In the local alignment, the majority of detected promoters are concentrated towards the end of the sequence, while in the statistical alignment, the promoter positions are more evenly distributed. This difference is primarily due to the fact that in local alignment algorithm, the last alignment with the highest score was considered as the correct one, introducing a bias towards later positions in the sequence. The outcome of the local alignment can also vary depending on the operator's choice of alignment strategy, such as selecting the first alignment or the one with the most consecutive "W"s. In contrast, the statistical alignment approach is less influenced by such decisions, making it appear more robust. However, it too is not entirely immune to operator influence, as factors like promoter length (set to 6 bases) and the predefined threshold (set to -6) also affect the results.

In conclusion, both algorithms effectively perform the task of promoter detection, each with its own strengths and limitations.