CSE4065- Computational Genomics

Programming Assignment 1

Group Members

Abdullah Enes Dizer -150119880

Yağmur Koçoğlu - 150119715

DNA String Mutation and Motif Finding Algorithms Report

1. Introduction

This report provides a detailed overview of DNA string mutation and motif finding algorithms implemented in Python. The implemented algorithms include Randomized Motif Search and Gibbs Sampler. The primary objective of these algorithms is to identify conserved motifs within a set of DNA sequences, which can provide insights into regulatory elements, protein binding sites, and other biologically significant regions.

2. Implemented Functions and Algorithms

2.1. DNA String Mutation

The **generate_dna_string(length)** function generates a random DNA string of a specified length. The **mutate_10mer(dna_string)** function mutates a randomly chosen 10-mer within a given DNA string. It introduces four random mutations within the 10-mer sequence.

2.2. Motif Finding Algorithms

2.2.1. Randomized Motif Search

The **randomized_motif_search(dna_strings, k)** function implements the Randomized Motif Search algorithm. It iteratively identifies motifs within a set of DNA strings by selecting random motifs, constructing a profile matrix, and updating the motifs based on the profile matrix until convergence. This algorithm aims to find the most probable motifs across the input DNA sequences.

2.2.2. Gibbs Sampler

The **gibbs_sampler(dna_strings, k, iterations)** function implements the Gibbs Sampler algorithm. It iteratively identifies motifs similar to Randomized Motif Search but utilizes a probabilistic approach to select motifs for updating. It randomly selects a motif from one of the input sequences, constructs a profile matrix excluding that motif, and updates the motif based on the profile matrix. This process repeats for a specified number of iterations to converge towards the optimal motifs.

2.3. Motif Scoring and Profiling

- The **score_motifs(motifs)** function calculates the score of a set of motifs based on their consensus sequence. It counts the number of mismatches between individual motifs and the consensus sequence.
- The **form_profile(motifs)** function constructs a profile matrix based on a set of motifs. The profile matrix represents the probabilities of each nucleotide at each position in the motifs.
- The **profile_most_probable_kmer(text, k, profile)** function identifies the most probable k-mer within a given DNA sequence based on a profile matrix. It calculates the probability of each k-mer and selects the one with the highest probability.

2.2.3. Median String Algorithm

The **median_string(dna_strings, k)** function implements the Median String Algorithm, which aims to identify the median motif among a set of DNA sequences. The median motif is defined as the k-mer motif that minimizes the total Hamming distance to all other motifs in the set.

The **hamming_distance(s1, s2)** function calculates the Hamming distance between two strings **s1** and **s2**. Hamming distance is defined as the number of positions at which corresponding symbols differ between two strings.

The algorithm iterates through all possible k-mers (4^k possibilities) and calculates the total Hamming distance between each k-mer and all motifs in the set of DNA strings. It selects the k-mer with the minimum total Hamming distance as the median motif.

- **Loop**: Iterate through all possible k-mers.
- **Pattern Generation**: Generate each k-mer pattern using base 4 representation (A, C, G, T).
- **Distance Calculation**: Calculate the total Hamming distance between the current k-mer and all motifs in the set of DNA strings.
- **Median Selection**: Update the minimum distance and the median motif if the calculated distance is less than the current minimum.

The function returns the identified median motif, which represents the most probable motif among the input DNA sequences.

RUN-1 / Randomized Motif Search – Gibbs Sampler Algorithms

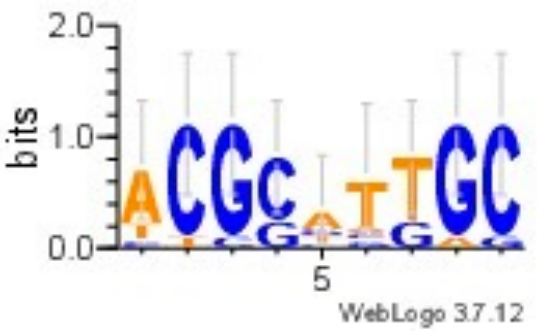
```
************************************************************
Running for k=9
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 21
Average score (Randomized Motif Search): 25.2
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 17
Average score (Gibbs Sampler): 20.6
------------------------------------------------------------
Randomized Motif Search - Best Motifs:
Motif 1: ACGGCTGGC
Motif 2: ACCCAGTGC
Motif 3: ACGGGTTGC
Motif 4: TCGCATTGG
Motif 5: ACGCTAGGC
Motif 6: TCGCATTGC
Motif 7: CCGCATTGC
Motif 8: ACGCTCTGC
Motif 9: ATGGGTTGC
Motif 10: ACGCATGAC
Consensus string (Randomized Motif Search): ACGCATTGC

Gibbs Sampler - Best Motifs:
Motif 1: CATGCGATG
Motif 2: CCCGAGTTG
Motif 3: GACGAAATG
Motif 4: CACGCGATG
Motif 5: AACGCGTTG
Motif 6: CTCGAGATG
Motif 7: CTCGAGATT
Motif 8: TACGAGACG
Motif 9: CACGAGATT
Motif 10: CATGAGATG
Consensus string (Gibbs Sampler): CACGAGATG
```

```
************************************************************
Running for k=10
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 31
Average score (Randomized Motif Search): 33.4
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 26
Average score (Gibbs Sampler): 26.4
------------------------------------------------------------
```

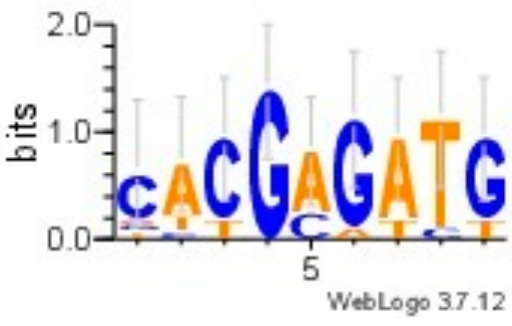
```
************************************************************
Running for k=11
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 33
Average score (Randomized Motif Search): 36.6
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 26
Average score (Gibbs Sampler): 29.2
------------------------------------------------------------
```

```
Running for k=9
Time (Randomized Motif Search): 0.12 seconds
Time (Gibbs Sampler): 10.00 seconds

Running for k=10
Time (Randomized Motif Search): 0.08 seconds
Time (Gibbs Sampler): 9.21 seconds

Running for k=11
Time (Randomized Motif Search): 0.11 seconds
Time (Gibbs Sampler): 9.19 seconds
```

Randomized Motif Search



Gibbs Sampler



RUN-2 / Randomized Motif Search – Gibbs Sampler Algorithms

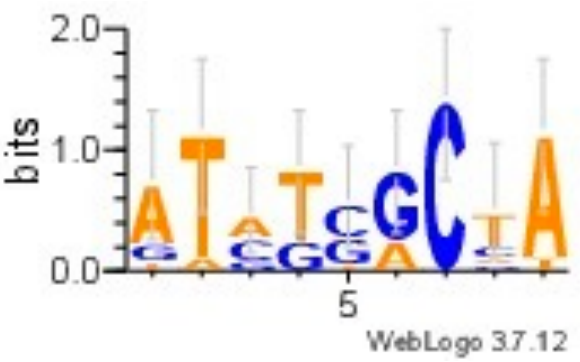
```
************************************************************
Running for k=9
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 26
Average score (Randomized Motif Search): 28.4
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 20
Average score (Gibbs Sampler): 22.4
------------------------------------------------------------
Randomized Motif Search - Best Motifs:
Motif 1: ATCTGGCTA
Motif 2: ATCGTGCCA
Motif 3: GTAGGGCTA
Motif 4: GTGTCGCCA
Motif 5: AAATGGCAA
Motif 6: ATGGCACTT
Motif 7: ATCTCACTA
Motif 8: ATATGACTA
Motif 9: ATATCGCGA
Motif 10: TTCTCGCTA
Consensus string (Randomized Motif Search): ATATCGCTA

Gibbs Sampler - Best Motifs:
Motif 1: GATGGTATT
Motif 2: GATAATATT
Motif 3: TATAACATT
Motif 4: GAGAAGATC
Motif 5: GATACTATT
Motif 6: GATCGTAGT
Motif 7: GTTAATATC
Motif 8: GATGGTATT
Motif 9: GATATAATT
Motif 10: GAAAACAAT
Consensus string (Gibbs Sampler): GATAATATT
```

```
************************************************************
Running for k=10
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 30
Average score (Randomized Motif Search): 32.8
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 22
Average score (Gibbs Sampler): 25.0
------------------------------------------------------------
```

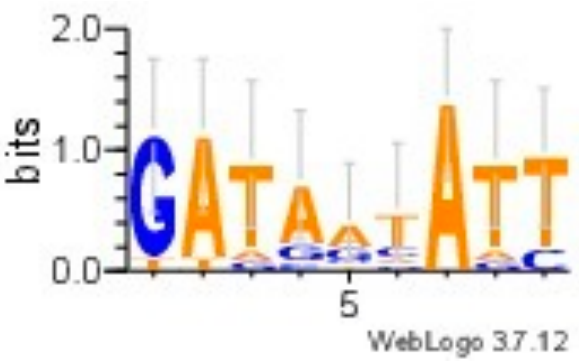
```
************************************************************
Running for k=11
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 33
Average score (Randomized Motif Search): 37.8
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 25
Average score (Gibbs Sampler): 30.6
------------------------------------------------------------
```

```
Running for k=9
Time (Randomized Motif Search): 0.09 seconds
Time (Gibbs Sampler): 8.86 seconds

Running for k=10
Time (Randomized Motif Search): 0.10 seconds
Time (Gibbs Sampler): 9.98 seconds

Running for k=11
Time (Randomized Motif Search): 0.11 seconds
Time (Gibbs Sampler): 10.79 seconds
```

Randomized Motif Search



Gibbs Sampler



RUN-3 / Randomized Motif Search – Gibbs Sampler Algorithms

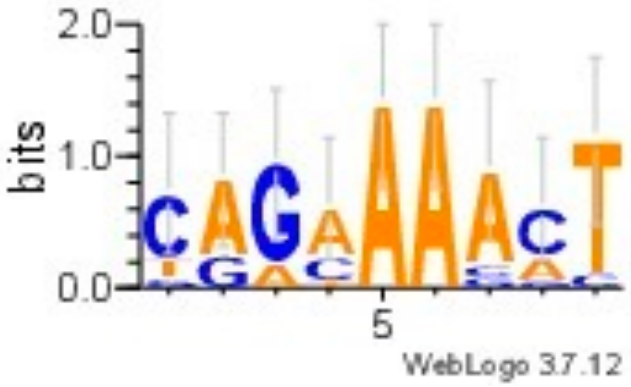
```
************************************************************
Running for k=9
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 19
Average score (Randomized Motif Search): 24.8
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 18
Average score (Gibbs Sampler): 21.2
------------------------------------------------------------
Randomized Motif Search - Best Motifs:
Motif 1: CAACAAACT
Motif 2: CAGAAAAAT
Motif 3: TAGAAAAAT
Motif 4: TAGAAAACT
Motif 5: CGGAAAACT
Motif 6: CAGAAAAAT
Motif 7: GAACAAACT
Motif 8: CGGAAACGT
Motif 9: CGGCAAGCC
Motif 10: CAGTAAACT
Consensus string (Randomized Motif Search): CAGAAAACT

Gibbs Sampler - Best Motifs:
Motif 1: ATGACCGCT
Motif 2: TTGACCTCT
Motif 3: TTGGCCGGT
Motif 4: TTGGCCGCT
Motif 5: TTGGCCCTC
Motif 6: TCGGCCACT
Motif 7: TTGGCACCG
Motif 8: ATGGCCGTT
Motif 9: GTTGCCGCT
Motif 10: TAGGCCGCT
Consensus string (Gibbs Sampler): TTGGCCGCT
```

```
************************************************************
Running for k=10
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 24
Average score (Randomized Motif Search): 29.4
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 23
Average score (Gibbs Sampler): 24.4
------------------------------------------------------------
```

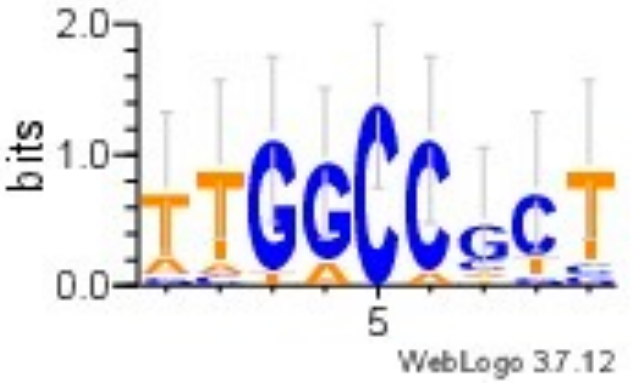
```
************************************************************
Running for k=11
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 35
Average score (Randomized Motif Search): 37.6
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 29
Average score (Gibbs Sampler): 29.4
------------------------------------------------------------
```

```
Running for k=9
Time (Randomized Motif Search): 0.13 seconds
Time (Gibbs Sampler): 9.30 seconds

Running for k=10
Time (Randomized Motif Search): 0.10 seconds
Time (Gibbs Sampler): 9.20 seconds

Running for k=11
Time (Randomized Motif Search): 0.11 seconds
Time (Gibbs Sampler): 10.83 seconds
```

Randomized Motif Search



Gibbs Sampler



RUN-4 / Randomized Motif Search – Gibbs Sampler Algorithms

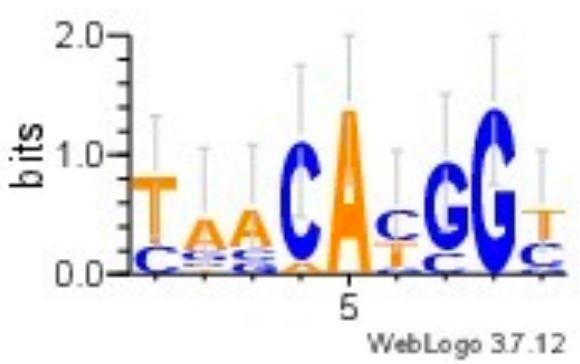
```
************************************************************
Running for k=9
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 24
Average score (Randomized Motif Search): 26.4
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 17
Average score (Gibbs Sampler): 21.6
------------------------------------------------------------
Randomized Motif Search - Best Motifs:
Motif 1: TAACACCGC
Motif 2: CAACATGGG
Motif 3: TAAAAGGGT
Motif 4: TTGCATGGT
Motif 5: TGACACGGC
Motif 6: CCACACCGT
Motif 7: TACCACGGT
Motif 8: CGCCACGGC
Motif 9: TAGCATGGC
Motif 10: TAACATGGT
Consensus string (Randomized Motif Search): TAACACGGT

Gibbs Sampler - Best Motifs:
Motif 1: TCATCTTCG
Motif 2: TCATCTACC
Motif 3: TCAGGTCCG
Motif 4: TCATGGACG
Motif 5: TTAACTACG
Motif 6: TCCTCTACG
Motif 7: TCATCAATG
Motif 8: TAATCCAAG
Motif 9: TCATCCACG
Motif 10: ACATCTACG
Consensus string (Gibbs Sampler): TCATCTACG
```

```
************************************************************
Running for k=10
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 32
Average score (Randomized Motif Search): 33.8
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 24
Average score (Gibbs Sampler): 24.8
------------------------------------------------------------
```

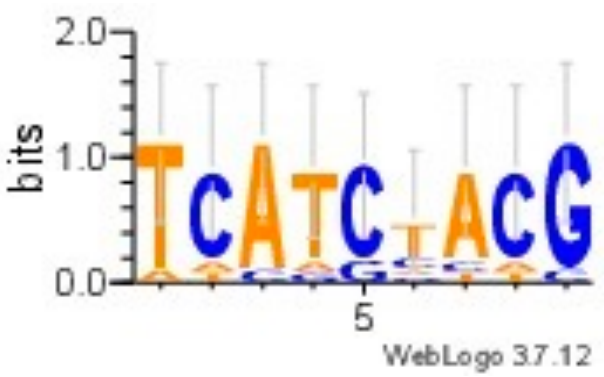
```
************************************************************
Running for k=11
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 32
Average score (Randomized Motif Search): 36.6
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 27
Average score (Gibbs Sampler): 32.4
------------------------------------------------------------
```

```
Running for k=9
Time (Randomized Motif Search): 0.07 seconds
Time (Gibbs Sampler): 9.15 seconds

Running for k=10
Time (Randomized Motif Search): 0.14 seconds
Time (Gibbs Sampler): 12.72 seconds

Running for k=11
Time (Randomized Motif Search): 0.13 seconds
Time (Gibbs Sampler): 12.10 seconds
```

Randomized Motif Search



Gibbs Sampler



RUN-5 / Randomized Motif Search – Gibbs Sampler Algorithms

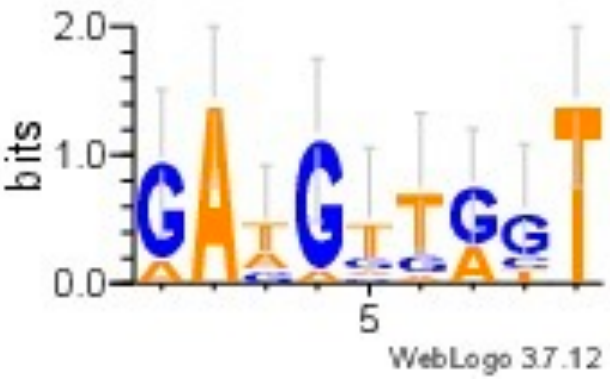
```
************************************************************
Running for k=9
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 23
Average score (Randomized Motif Search): 25.2
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 19
Average score (Gibbs Sampler): 22.6
------------------------------------------------------------
Randomized Motif Search - Best Motifs:
Motif 1: GATATTAGT
Motif 2: GAAGGTATT
Motif 3: GATGATGGT
Motif 4: GAAGGGGCT
Motif 5: GATGTTACT
Motif 6: GATGCTGGT
Motif 7: GAGGTTGGT
Motif 8: AATGTTGTT
Motif 9: AAGGTGAGT
Motif 10: GAAGTAGGT
Consensus string (Randomized Motif Search): GATGTTGGT

Gibbs Sampler - Best Motifs:
Motif 1: AGCCTTATC
Motif 2: TACCTTAAT
Motif 3: TACCTTAAC
Motif 4: TACCTTCTT
Motif 5: GACCTTCGC
Motif 6: CCCCTTATA
Motif 7: TCCCTTATT
Motif 8: TACCCTATG
Motif 9: TACCTCATC
Motif 10: TACCTTATA
Consensus string (Gibbs Sampler): TACCTTATC
```

```
************************************************************
Running for k=10
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 26
Average score (Randomized Motif Search): 30.2
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 23
Average score (Gibbs Sampler): 24.8
------------------------------------------------------------
```

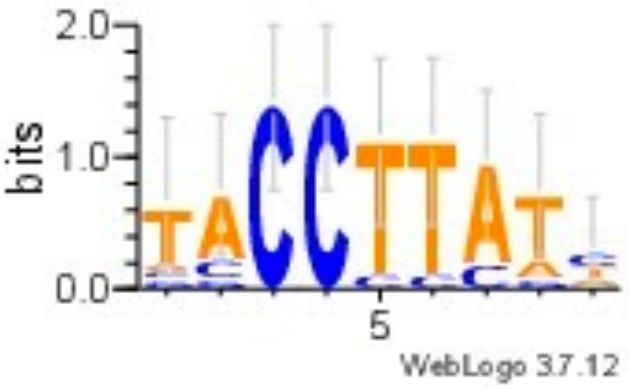
```
************************************************************
Running for k=11
------------------------------------------------------------
*Randomized Motif Search*
Best score (Randomized Motif Search): 27
Average score (Randomized Motif Search): 34.6
------------------------------------------------------------
*Gibbs Sampler*
Best score (Gibbs Sampler): 26
Average score (Gibbs Sampler): 29.2
------------------------------------------------------------
```

```
Running for k=9
Time (Randomized Motif Search): 0.08 seconds
Time (Gibbs Sampler): 11.92 seconds

Running for k=10
Time (Randomized Motif Search): 0.14 seconds
Time (Gibbs Sampler): 11.04 seconds

Running for k=11
Time (Randomized Motif Search): 0.17 seconds
Time (Gibbs Sampler): 11.85 seconds
```

Randomized Motif Search



Gibbs Sampler



Median String Algorithm

```
**********
Running Median String for k=9
Median String: GGGGGTCGG
Time (Median String): 1350.92 seconds
**********
Running Median String for k=10
Median String: GCCGGTTGCC
Time (Median String): 6530.48 seconds
**********
Running Median String for k=11
```

When we ran the median string algorithm, we had to terminate the programme because k=11 did not give any output despite running for more than 24 hours.

Conclusion

Comparing the performance of the three algorithms employed in this project reveals notable trends in their behavior. Notably, as the value of k increases, both the Randomized Motif Search and Gibbs Sampler algorithms yield higher scores. Regarding runtime, it is evident that Gibbs Sampler and the Median String algorithm exhibit longer processing times compared to Randomized Motif Search. Particularly, as k increases, the Median String algorithm experiences a significant increase in processing time. Consequently, it appears to be a less efficient algorithm relative to the others. Conversely, Randomized Motif Search consistently delivers comparable scores to Gibbs Sampler but within significantly shorter timeframes. Ultimately, Gibbs Sampler stands out for its ability to explore all possible permutations, resulting in superior outcomes relative to the other algorithms.