Motif Detection and Entropy-Based Filtering in DNA Sequences

Manuel Ricardo Guerrero Cuéllar
Department of System Engineering
University Francisco José de Caldas
Email: mrguerreroc@udistrital.edu.co

Abstract—**This report addresses the challenge of detecting motifs—frequent patterns within DNA sequences—and filtering them using entropy-based techniques. The system comprises two main components: dataset generation and motif detection with entropy filtering. The project focuses on the computational complexity and performance analysis of these tasks and offers experimental results on varying entropy thresholds and motif sizes.**

I. Systemic Analysis

This project tackles the problem of detecting motifs in DNA sequences and filtering them based on Shannon entropy. The system is divided into two primary components:

A. Dataset Generation

Artificial DNA sequence datasets were generated with predefined nucleotide probabilities. Each sequence consists of a random arrangement of the four nucleotide bases (A, C, G, T), controlled by the parameters: number of sequences (n), sequence length (m), and nucleotide probabilities.

B. Motif Detection and Entropy Filtering

The detection and filtering process includes two steps:

1) Motif Detection: An algorithm detects recurring motifs of nucleotide base combinations of size s ($4 \leq s \leq 10$).

2) Entropy-Based Filtering: Shannon entropy is used to filter out sequences with excessive repetition to retain more chaotic sequences. Entropy thresholds between 0.5 and 2.0 were used to control the diversity of the dataset.

The system handles large datasets efficiently, using parallel processing for optimized performance and modular architecture for scalability, supporting datasets ranging from 1,000 to 2 million sequences.

II. Complexity Analysis

The computational complexity is divided into:

A. Motif Detection

The time complexity is $O(n \cdot m)$, where n is the number of sequences, and m is the length of each sequence. The algorithm scans for all possible motif combinations of size s across all sequences.

B. Entropy Calculation

Entropy calculation also has a time complexity of $O(n \cdot m)$, requiring counting base occurrences and computing entropy using Shannon's formula.

To reduce computation time, multithreading was implemented using Java's ExecutorService, which distributes tasks across multiple threads. This approach mitigates the exponential growth in computation time for large datasets, especially when applying higher entropy thresholds.

III. Chaos Analysis

Chaos is tied to the randomness in DNA sequences, measured by Shannon entropy. The goal is to retain diverse sequences by filtering out repetitive ones.

A. Entropy Thresholding

Entropy thresholds significantly affect the filtering process. Lower thresholds (e.g., 0.5) allow more repetitive sequences, while higher thresholds (e.g., 2.0) filter out these sequences, leaving more chaotic and diverse ones. Experimentation was used to determine the ideal entropy threshold, balancing data diversity and motif detection.

IV. Results

Two sets of results were generated by varying entropy thresholds and applying the ideal entropy filter:

A. Results from Entropy Variations

Dataset Size ($n = 1000$):

- Motif size 5: Frequent motifs include “GCCTT”, “CGCAG”, and “GGAAT”, with detection times between 8 ms and 80 ms, depending on the entropy threshold.
- Motif size 10: Detected motifs such as “TAATAAATGC” and “TGACGTTTTC”, with entropy values 0.5, 1.0, and 1.5.

Dataset Size ($n = 500,000$):

- Motif “GACCA” detected with an entropy threshold of 0.5 took 1181 ms, while “CGCTA” took 1498 ms.

B. *Results from Ideal Entropy*

Dataset Size ($n = 1000$):

- Motif size 5: Motifs such as "GAACG" and "CTAGT" detected in 4-5 ms.
- Motif size 10: "CGAGTACCGT" detected in 2 ms.

Dataset Size ($n = 500,000$):

- Motifs like "ATCGG" and "AGCCCATGCG" took 815-1986 ms to be detected.

Results are stored in two CSV files:

- *Motif_Data_Entropy_Variations.csv*
- *Motif_Data_Ideal_Entropy.csv*

V. Discussion of Results

A. *Results from Ideal Entropy*

The following table presents the experimental results obtained using the ideal entropy threshold (1.0), showing motif detection times for different dataset sizes and motif sizes:

TABLE I
Results from Ideal Entropy Threshold (1.0)

Database Size	**Probability of Bases**	**Motif Size**	**Motif**	**Motif Occurrences**	**Time to Find Motif**	**Entropy**
1000	0.25, 0.25, 0.25, 0.25	5	GAACG	1	4 ms	1.0
1000	0.25, 0.25, 0.25, 0.25	10	CGAGTACCGT	1	2 ms	1.0
1000	0.25, 0.25, 0.25, 0.25	5	CTAGT	1	5 ms	1.0
1000	0.25, 0.25, 0.25, 0.25	10	CGCCATTCGG	1	10 ms	1.0
500000	0.25, 0.25, 0.25, 0.25	5	ATCGG	1	815 ms	1.0
500000	0.25, 0.25, 0.25, 0.25	10	AGCCCATGCG	1	767 ms	1.0
500000	0.25, 0.25, 0.25, 0.25	5	ATATA	1	1986 ms	1.0
500000	0.25, 0.25, 0.25, 0.25	10	GTCGGTTATG	1	2511 ms	1.0

B. *Entropy Filtering*

Lower thresholds allow more repetitive sequences and more frequent motifs, while higher thresholds produce fewer, more diverse motifs.

C. *Motif Size*

Shorter motifs (e.g., 5 bases) appear more frequently but may be biologically less significant. Longer motifs (e.g., 10 bases) are harder to detect but often carry more biological relevance.

D. *Performance*

Larger datasets result in higher computation times, mitigated by parallel processing. Further optimization is needed for extremely large datasets.

VI. Conclusions

This project implemented a scalable system for generating DNA sequences, detecting motifs, and filtering sequences using Shannon entropy. Key conclusions are:

- Entropy filtering effectively retains diverse, chaotic sequences with biologically significant motifs.
- Motif size and entropy thresholds are crucial for motif detection, providing flexibility for different analysis environments.