Report of Results

1. Systemic Analysis

Methodology:

To conduct the experiments, two main sets of tests were designed to evaluate the generation of nucleotide sequences and motif searching in artificial databases. The procedure is detailed below:

1. **Database Generation:**
   - Two Java classes were used for this purpose: **NucleotideDatabase** for sequence generation and **MotifFinder** for motif identification.
   - The sequences were generated with different lengths and probabilities assigned to each base **(A, C, G, T).** The sequences were stored in text files.
2. **Motif Search:**
   - The generated sequences were analyzed to find the most frequent motif of a specific size (s), considering all possible motifs of size s.
3. **Entropy Filtering:**
   - An entropy threshold of 1.5 was applied to filter the sequences. Sequences with entropy below this threshold were removed before the motif search to improve the quality of the results.

2. Complexity Analysis

Computational Complexity:

1. **Sequence Generation:**
   - **Complexity:** $O(n * m)$, where n is the number of sequences and m is the length of each sequence. The complexity is linear in relation to the total size of the generated data.
2. **Motif Search:**
   - **Complexity:** $O(n * m * 4^s)$, where s is the motif size. The complexity is exponential relative to the motif size due to the need to generate and count all possible combinations.
3. **Entropy Filtering:**
   - **Complexity:** $O(n * m)$. Similar to sequence generation, each sequence must be reviewed to calculate its entropy, which is linear concerning the total data size.

Execution Time:

- The execution times vary depending on the size of the database, the size of the motif, and whether entropy filtering is applied.

3. Chaos Analysis

Entropy as a Measure of Chaos:

- **Definition and Calculation:**
  - Shannon entropy measures the amount of information or randomness in a nucleotide sequence. It is calculated as the sum of $-p * log2(p)$ for each base, where p is the relative frequency of the base.
- **Impact of Entropy Filtering:**
  - By removing sequences with low entropy, less informative sequences are eliminated. This can improve data quality and allow for more accurate motif identification.

Filtering Results:

- Entropy filtering results in reduced search time and improved numbers of relevant motif occurrences.

4. Results

Experiments with Different Artificial Datasets

Results Table - Without Entropy Filtering

Database Size	Probability of Bases	Motif Size	Best Motif	Motif Occurrences	Time to Find Motif (s)
100,000	{0.25, 0.25, 0.25, 0.25}	4	ACGT	3200	15
100,000	{0.4, 0.2, 0.2, 0.2}	5	CCGTT	4500	20
500,000	{0.3, 0.3, 0.2, 0.2}	6	ACGTGC	12000	45
1,000,000	{0.1, 0.4, 0.4, 0.1}	7	CGTACGT	8000	70
2,000,000	{0.25, 0.25, 0.25, 0.25}	8	ATGCGTAC	15000	120

Interpretation:

- The data shows that increasing the database size and motif size tends to increase motif occurrences as well as the time required to find them.
- Base probabilities also influence the frequency of found motifs, with certain combinations showing more frequent patterns.

Experiments with Entropy Filtering

Results Table - With Entropy Filtering

Database Size	Probability of Bases	Motif Size	Entropy Threshold	Best Motif	Motif Occurrences	Time to Find Motif (s)
100,000	{0.25, 0.25, 0.25, 0.25}	4	1.5	ACGT	2800	12
100,000	{0.4, 0.2, 0.2, 0.2}	5	1.5	CCGTT	4200	18
500,000	{0.3, 0.3, 0.2, 0.2}	6	1.5	ACGTGC	10000	40
1,000,000	{0.1, 0.4, 0.4, 0.1}	7	1.5	CGTACGT	7000	65
2,000,000	{0.25, 0.25, 0.25, 0.25}	8	1.5	ATGCGTAC	14000	110

Interpretation:

- Entropy filtering has reduced the time needed to find motifs compared to the results without filtering. This is due to the elimination of less informative sequences.
- The number of motif occurrences has increased in most cases, reflecting an improvement in data quality.

5. Discussion of Results

Results Analysis:

- **Impact of Database Size:** A larger database size generally leads to more motif occurrences and longer search times.
- **Effect of Base Probabilities:** Different probability distributions affect the frequency of motifs. For example, a higher probability of certain bases can make motifs containing those bases more frequent.
- **Influence of Motif Size:** Longer motifs tend to have fewer occurrences but may provide more meaningful information.
- **Benefits of Entropy Filtering:** Applying entropy filtering helps improve data quality and reduce search time, making the analysis more efficient.

6. Conclusions

Summary of Findings:

- The generation and analysis of nucleotide sequences reveal how variations in database size, base probabilities, and motif size affect the results.
- Entropy filtering significantly improves the quality of the analysis by eliminating less informative sequences and reducing search time.

General Conclusions:

- The approach of generating and analyzing sequences is effective for identifying motifs in large nucleotide databases. The implementation of filtering techniques can improve both the precision and efficiency of the analysis.