Name: Jaider Santiago Avila Robles - 20231020200

Systemic Analysis

The goal of this project was to create an artificial database of genetic sequences and identify repeated motifs using different base probabilities and sequence sizes. The system was divided into three main steps:

1. **Database Creation:** A "database" of genetic sequences was generated using different probabilities for the letters A, C, G, and T. This process utilized threads to efficiently generate the sequences and store them in a file.
2. **Motif Discovery:** The most frequent motifs were identified by analyzing each sequence in the database. The motif size (between 4 and 10) was passed as a parameter, and the system searched for all possible combinations within the specified motif length.
3. **Entropy Filter:** A Shannon entropy filter was applied to remove sequences with excessive base repetitions. A threshold of 1.5 was used to define a "chaotic" sequence. This reduced the number of sequences analyzed for motif discovery.

Complexity Analysis

The system's complexity is mainly driven by three components:

- **Sequence Generation:** The sequence generation process is O(n·m)O(n \cdot m)O(n·m), where nnn is the number of sequences and mmm is the length of each sequence.
- **Motif Searching:** Finding motifs involves iterating through all substrings of length sss (motif size) for each sequence, which gives a complexity of O(n·(m−s+1))O(n \cdot (m-s+1))O(n·(m−s+1)). This step becomes computationally expensive as the number of sequences increases.
- **Entropy Filtering:** The calculation of entropy is O(m)O(m)O(m) per sequence, as each sequence is processed to determine its level of chaos (randomness). Filtering the sequences before motif searching reduces the total number of sequences to analyze, leading to more efficient motif discovery.

Chaos Analysis

In this experiment, sequences with entropy below 1.5 were filtered, leaving only the more chaotic sequences for motif analysis. The entropy filter helps to prevent repeated motifs from dominating the search process and ensures a more diverse dataset for pattern discovery.

Results

The results of both experiments (with and without the entropy filter) are presented in the following tables:

- **Without Entropy Filter:**

Tamaño de la base de datos m = 10	Probabilidad de bases(A, C, G, T)	Tamaño del motif	Ocurrencias del motif	Tiempo para encontrar el motif (ms)
10000	0.25, 0.25, 0.25, 0.25	5	12(AATAC)	12
100000	0.4, 0.2, 0.2, 0.2	5	612(AAAAA)	52
20000	0.25, 0.25, 0.25, 0.25	6	9(TGAATA)	24
50000	0.3, 0.3, 0.2, 0.2	4	307(CACC)	61

- **With Entropy Filter, Threshold 1.5:**

Tamaño de la base de datos	Probabilidad de bases(A, C, G, T)	Tamaño del motif	Ocurrencias del motif	Tiempo para encontrar el motif (ms)
10000	0.25, 0.25, 0.25, 0.25	5	12(AATAC)	12
100000	0.4, 0.2, 0.2, 0.2	5	585(AAAAA)	80
20000	0.25, 0.25, 0.25, 0.25	6	11(CGCAGC)	24
50000	0.3, 0.3, 0.2, 0.2	4	309(AACC)	53

Discussion of Results

The main differences between the two experiments are related to the effects of the entropy filter:

- **Impact on Occurrences:** Generally, after applying the entropy filter, the occurrences of motifs slightly decreased, except in the dataset of 50,000 sequences, where they increased. This suggests that the entropy filter effectively removed some highly repetitive sequences, leading to a more balanced set of results.

- **Execution Time:** The time required to find motifs generally increased when the entropy filter was applied. This is because the remaining sequences after filtering are more chaotic and complex, requiring more time to analyze.
- **Variety of Motifs:** The motifs discovered with and without the entropy filter are mostly the same; however, in the 20,000 sequence dataset, the motif changed from "TGAATA" to "CGCAGC", indicating that the filter allows for the detection of different patterns in more chaotic data.

Conclusions

Applying an entropy filter based on Shannon entropy helped reduce sequence repetition and provided a clearer view of motif distribution in chaotic datasets. However, this comes at the cost of increased computational time due to the complexity of the remaining sequences. The entropy threshold should be chosen carefully, depending on the desired balance between randomness and repetition in the dataset.