UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Systems Analysis

Carlos Andrés Sierra

Workshop 1 Entropy and divide and conquer

Barrera PulidoSamuel Andrés **Código: 20232020156**

Bogotá, D.C.

15 de septiembre del 2024.

1. **Systemic Analysis**

To approach this bioinformatics project, we began by generating an artificial database of genetic sequences, where each sequence is composed of the typical DNA bases: A, C, G, and T. I implemented a concurrent approach to generate between 1,000 and 2 million sequences with variable lengths ranging from 5 to 100 characters. Each base is selected using user-defined probabilities, ensuring that these sum to 1 to avoid inconsistencies in the data.

The concurrent approach allows the task of sequence generation to be divided among multiple execution threads, speeding up the process and optimizing system resource usage. Additionally, the generated sequences are stored in a `database.txt` file, (which contains the tests logged in the project's folder), facilitating further processing and analysis.

In terms of task division, the process follows a divide-and-conquer strategy. Each sequence is independently generated in different threads, contributing to system scalability and allowing large volumes of data to be processed in parallel. This approach is critical for efficiently managing sequences when their number increases significantly.

2. **Algorithmic Complexity**

To find motifs in the sequences, I implemented an algorithm that searches for patterns of size s (where $4 \leq s \leq 10$) in each sequence. The motif search uses a direct search algorithm that scans each sequence for the desired pattern. If the motif is repeatedly found in a sequence, the number of occurrences is counted and compared with other sequences.

The search algorithm has a complexity of O(n*m), where n is the number of sequences and m is the length of each sequence. Although all possible combinations of nucleotides for motifs of size s are not explored, search time is optimized by focusing only on the generated sequences and the specific motif input by the user.

Furthermore, by working concurrently, the motif search is distributed across different processing cores, reducing the overall time required to process large volumes of sequences.

3. **Chaos Analysis**

To ensure that the sequences are as diverse as possible and avoid excessive repetition of the same bases, Shannon Entropy is used as a measure of chaos in the system. Entropy evaluates the distribution of the bases A, C, G, and T in each sequence, where high entropy indicates a more diverse and chaotic sequence, and low entropy signals that one base repeats more frequently.

In the code, entropy is calculated for each generated sequence. Sequences with very low entropy can be discarded or marked as less useful for further analysis, as they contain less significant information. This ensures that the analyzed sequences maintain high variability, which is key in bioinformatics studies.

To understand the results obtained from Shannon Entropy, it is important to note that entropy is measured on a scale from 0 to 2, as it uses the logarithm of the number of elements in

base 2. Thus, the maximum value for this entropy analysis will be 2. A threshold between 0 and 1 indicates low entropy; from 1 to 1.5, it is considered medium entropy; and from 1.5 to 2, it is high entropy. For this first workshop, we need an entropy close to 2 for it to be a good chaotic system and not overly predictable.

4. **Results:**

Database Size	Motif Size & Motif Searched	Time to Find It	Shannon Entropy	Motifs Found
1000	4 (ATTC)	1.3s	1.999	10
2700	5(ACTGC)	3.71s	1.998	8
3200	5(ATACT)	3.579s	1.999	11
10000	4(CTGA)	15.743s	1.999	60
100000	5(GTGAC)	241.19s	1.999	133

5. **Discussion of Results**:

The entropy filter proved to be effective in enhancing the quality of sequences in motif analysis. By eliminating sequences with low entropy, the impact of redundant and repetitive sequences was reduced, leading to a more accurate and useful analysis. Shannon Entropy provided an appropriate measure of diversity in the sequences, helping to select those that best represented genetic variability.

The implementation of the entropy filter has proven to be an effective strategy for improving data quality in genetic sequence analysis. By selecting sequences with high entropy, a more diverse and representative dataset was achieved, resulting in better motif detection and greater accuracy in analysis. It is recommended to continue using the entropy filter in future studies and adjust the threshold as necessary to optimize results.

In general terms of the project, we can observe that for very large datasets, it tends to be somewhat slow, but it is indeed efficient since it is comparing millions of data points. Overall, it maintains a high entropy threshold, which is advantageous given that the user needs to specify the frequency of each letter on a scale from 0 to 1 (using commas instead of periods to separate decimals). Thus, the work was successfully completed and meets the requirements.

6. Conclusions:

Effectiveness of Concurrent Processing: The concurrent approach to generating sequences significantly reduces processing time, allowing the handling of large datasets efficiently. By distributing the workload across multiple threads, the system can generate sequences faster and utilize resources more effectively. This approach is essential for projects dealing with large volumes of genetic data.

Motif Detection Algorithm: The implemented direct search algorithm for motif detection is straightforward and effective for finding patterns of various sizes within sequences. The algorithm's complexity of O(n*m) is manageable for the dataset sizes tested, and its performance is enhanced by concurrent execution. However, for even larger datasets or more complex motif search requirements, further optimizations or alternative algorithms may be considered.

Shannon Entropy as a Quality Filter: The use of Shannon Entropy to filter sequences based on diversity proves to be a valuable technique. High-entropy sequences indicate greater genetic diversity and reduce the impact of repetitive data, improving the quality of the analysis. The results show that sequences with high entropy are more informative and useful for motif detection. Future work should involve refining the entropy threshold and exploring additional methods to enhance data quality further.

Overall Performance: The system effectively handles the generation and analysis of large datasets, maintaining high entropy and providing accurate motif detection. While the processing time increases with dataset size, the system's efficiency and the quality of results justify the approach. The project successfully meets the requirements and demonstrates the effectiveness of the implemented methods.