UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

FACULTAD DE INGENIERÍA

REPORT WORKSHOP 1
SYSTEMS ANALYSIS

Cesar Augusto Pulido Cuervo - 20222020048

Bogotá, 2024

Systemic Analysis

The project's Systemic Analysis is centered on finding recurrent sequences in a text file-stored database of randomly generated DNA sequences. The research creates sequences made up of nucleotides (A, C, G, and T) to mimic a biological database. Then, it uses computational methods to look for particular themes in these sequences. Utilizing effective distributed algorithms for pattern identification and string matching, the system is able to search through enormous datasets, count the number of times motifs appear, and examine their distribution.

Complexity analysis

The usage of hashmaps to store and retrieve motif occurrences and the concurrent scanning of the text file using several Java threads are the two main processes that are the subject of the project's complexity analysis. Because hash-based lookups take constant time, hashmap operations like inserting and accessing entries often have an average time complexity of O(1). In the worst situation, though, this can deteriorate to O(n) in collision situations. The I/O operations, which are usually O(n), where n is the file size, dominate the reading time complexity when reading a big text file concurrently with numerous threads. While using many threads promotes parallelism and may increase efficiency for CPU-bound operations, it also adds overhead for thread management and synchronization, which may have an adverse effect on performance.

Chaos Analysis

The produced DNA sequences' degree of unpredictability or randomness is measured using the Shannon entropy method. We quantify the information content of the sequence set by examining the database's motif occurrence probabilities.

$$H(X) = -\sum_{i} P(x_i) \log P(x_i)$$

The Shannon entropy measure sheds light on how equally dispersed the motifs are; larger values represent a more structured or repeated pattern, while lower values indicate more unpredictability and a more chaotic distribution of motifs. We can determine the degree of disorder in the nucleotide sequences and evaluate the complexity of the dataset by calculating the probability of motif occurrences. Finding patterns and abnormalities using this type of chaos analysis is essential as it may have ramifications for domains like genetic research and bioinformatics.

Results

With experiments generating different artificial datasets using sequence of size 10 and different probabilities for each base and different amount of sequences, and search motifs with different size changing the following parameters: Database Size, Probability of Bases and Motif Size.

Database size	Probability of bases	Motif size	Entropy
1000	0.25 - 0.25 - 0.25 - 0.25	4	7.8941
2000	0.1 - 0.5 - 0.2 - 0.2	4	6.9818
2000	0.3 - 0.1 - 0.1 - 0.5	3	5.0250
1500	0.25 - 0.25 - 0.25 - 0.25	2	3.9375

Discussion of results

A comparison of Shannon entropy based on different database sizes, nucleotide probabilities, and motif sizes is shown in the Results table. Here's a thorough explanation:

- An overall decrease in entropy values is observed as the database size grows (from 1000 to 2000). A drop in entropy may suggest a decrease in randomness or a more structured motif occurrence; a larger dataset offers a more thorough depiction of motif distribution. But in this changes is not more important than the other parameters
- The entropy in rows 1 and 4, where the probabilities of bases are equally distributed (0.25 for each base) is very different probably by the motif size in each one of the experiments.
- The table shows a strong correlation between base probabilities, motif size, and Shannon entropy-measured sequence randomization. More motif sizes and uniform base probabilities increase entropy, which indicates increased uncertainty in the motif distribution.

Conclusions

- This exercise helps to understand in a better way the entropy applied to a genetic system.
- Entropy is strongly influenced by variations in motif sizes, while skewed nucleotide distributions lead to reduced entropy and more predictable motif patterns.
- Greater entropy is correlated with larger motif sizes, suggesting that longer motifs add complexity to the sequence and increase the difficulty of predicting occurrences. This may be a crucial consideration when examining more intricate genetic patterns.