UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Report Workshop 1

Juan Sebastian Bedoya Serna

2023102029

Faculty of engineering; Universidad Distrital Francisco Jose De Caldas

System’s Analysis

Carlos Andrés Sierra Virgüez

September 14 – 2024

INTRODUCTION

The following document aims to provide clarity on the concepts of systemic analysis, complexity analysis, chaos analysis, results, discussion of results focused on the project carried out to find the motif of DNA sequences

SYSTEMIC ANALYSIS

The provided code implements a system to generate random DNA sequences, identify frequent motifs, apply Shannon Entropy to filter sequences, and conduct performance experiments. The system is composed of:

1. **Database Generation:** The project can generate a database of random DNA sequences based on user-specified lengths and probabilities for each base (‘A’,’C’.’G’,’T’).

2. **Motif Finding:** The project can read sequences from the generated database and finds the most frequent motifs of a given length within the sequences.

3. **Experimentation / Test Cases:** The project automates the running of experiments or distinct test cases by varying the parameters like database size, sequence length, and motif length. The results, including the motifs found, their counts, and the time taken, are saved in a CSV file

This system basically is designed to study different motif patterns in DNA sequences under different conditions.

In addition, the following table shows how the system works in terms of its inputs, processes and outputs:

ELEMENT	INPUTS	PROCESSES	OUTPUTS
Data Base Generation	• User-specified lengths • Probabilities for DNA bases (‘A’, ‘C’, ‘G’, ‘T’).	Generate random DNA sequences according to user specifications.	Random DNA sequences generated stored in a .TXT file.

Motif Finding	• DNA sequences from the file.	Identify most frequent motifs of a given length.	Frequent motifs found within sequences.
Experimentation	• Parameters such as the number of sequences, sequence length and motif length.	Conduct experiments by varying parameters and testing motifs found.	Results saved in CSV files.
Shannon Entropy Calculation	• DNA sequences from the file	Detect the entropy among all the sequences	The amount of entropy according to the Shannon model

On the other hand, this graph explains the elements of the system and their relations.

Elements

1. File Writing
2. Sequence generation
3. Motif finding
4. Shannon entropy calculation
5. Experiments

Relations

1. Motif finding uses the sequence generated to detect the motif of the sequences
2. Shannon entropy calculation use the sequence generated to calculate the entropy that exist in all the sequences.
3. Sequence generation show its results with the file writing in a .txt file
4. Motif finding show its results with file writing in a .csv file
5. Experiments use multiple time the results of the sequence generation to show different results
6. Experiments use motif finding to detect the motif in all iterations.
7. Experiments show its results with file writing in a .csv file

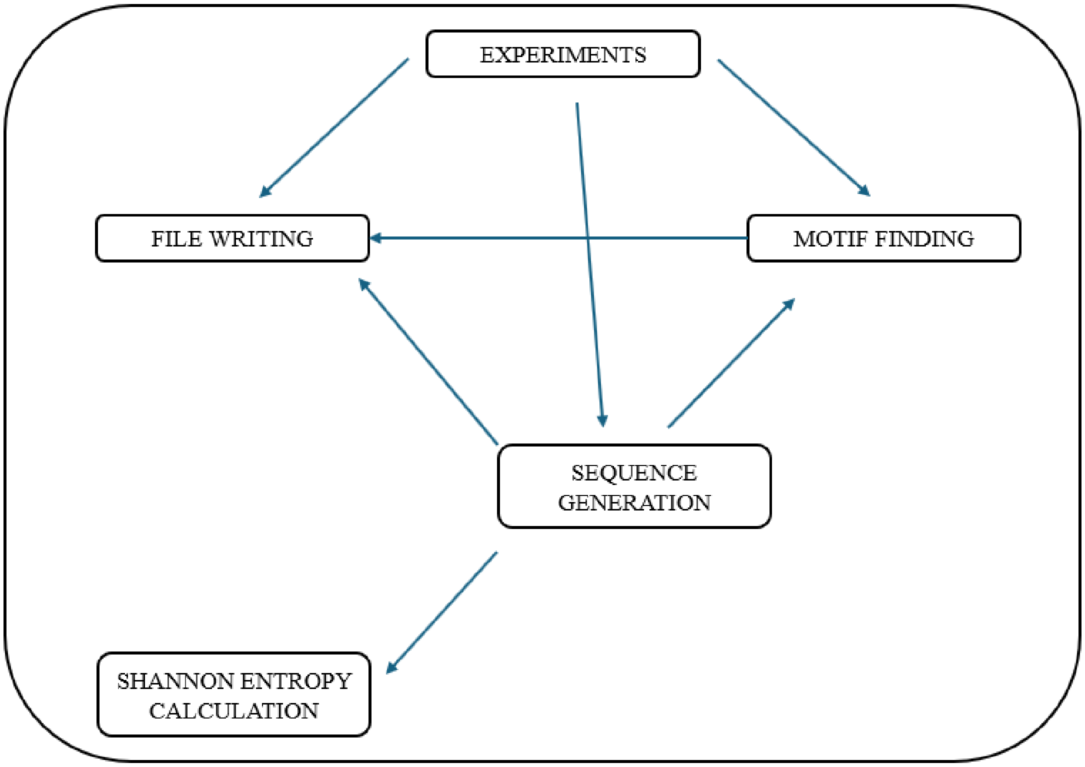


Image 1: Elements and their relations

COMPLEXITY ANALYSIS

To perform the project complexity analysis, two components will be considered: time and space.

Certainly, the time it will take for the program to complete all its processes will depend on the values entered by the user, the larger the value of n (Number of sequences), the larger the value of m (Length of sequences) or the length of the motif for the program to take more or less than expected.

Furthermore, finding motifs involves scanning through each sequence and generating all possible substrings of the length given.

On the other hand, the space occupied by the sequences could be said to be given by the multiplication n*m, so the greater the value of some of these two elements, the greater the space occupied by what is generated by the program.

CHAOS ANALYSIS

One of the elements that could generate more chaos in the project is the randomness of some data such as:

1. **Random DNA Generation:** The random selection of bases introduces variability in the generated sequences, generating chaos by presenting the uncertainty of not knowing at any time what the generated sequence will be, that also can lead to varied motif occurrences across different runs.

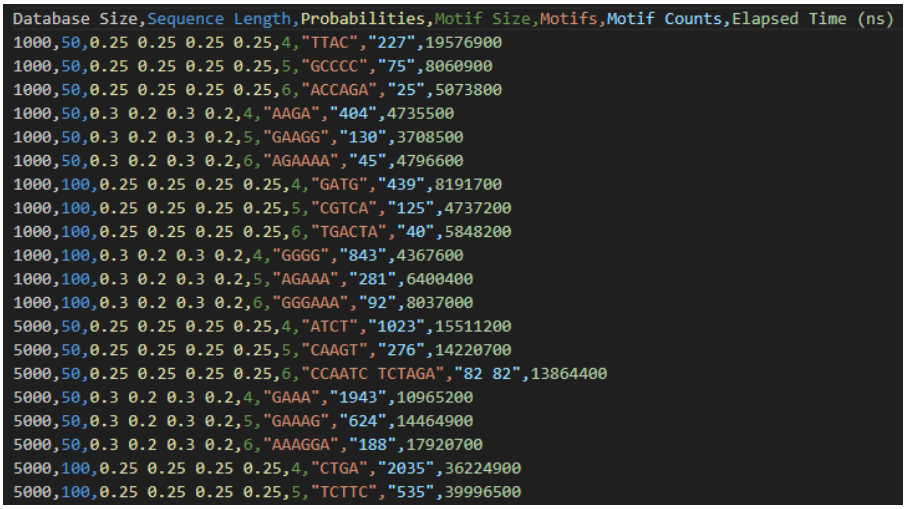
2. **Random Motif Length:** During motif finding, the length of motifs to be identified is randomly selected, adding another variable to consider. This randomness can affect the frequence, type and the time that take to detect motifs in each run.

Even though the system is designed to handle deterministic inputs, the randomness in the DNA sequence generation and motif length introduces an element of chaos, leading to unpredictable results.

RESULTS

Varying parameters such as database size, sequence length, and motif size, the results for several experiments were saved to a CSV file, showing the following trends:

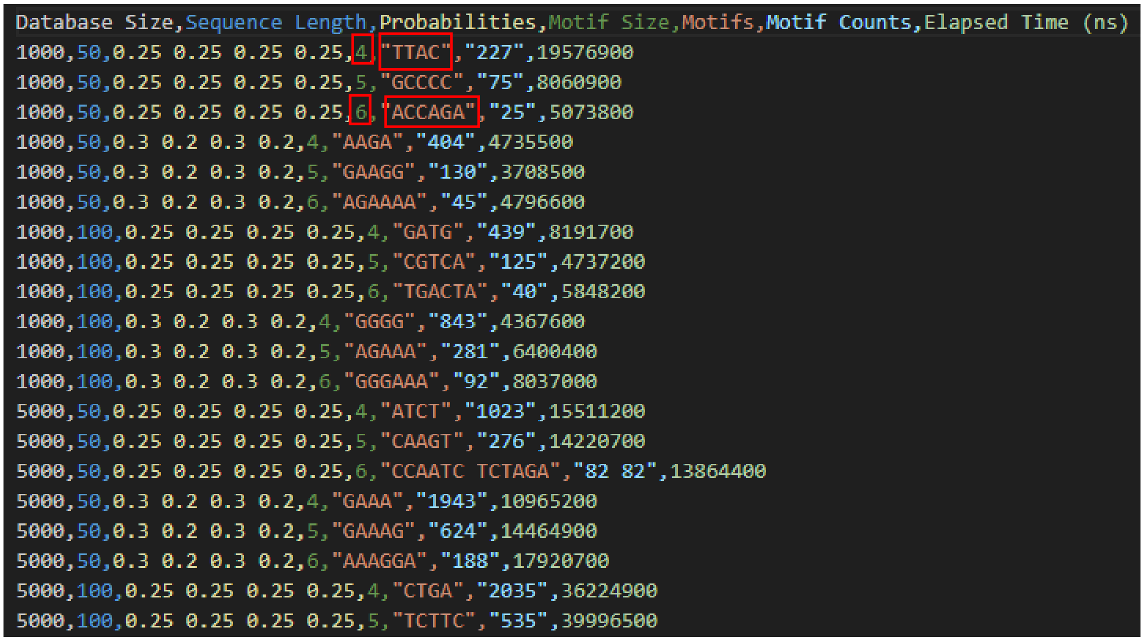
1. **Larger Databases:** As the size of the database (n) increases, the time required to find motifs increases significantly.



```
Database Size,Sequence Length,Probabilities,Motif Size,Motifs,Motif Counts,Elapsed Time (ns)
1000,50,0.25 0.25 0.25 0.25,4,"TTAC","227",19576900
1000,50,0.25 0.25 0.25 0.25,5,"GCCCC","75",8060900
1000,50,0.25 0.25 0.25 0.25,6,"ACCAGA","25",5073800
1000,50,0.3 0.2 0.3 0.2,4,"AAGA","404",4735500
1000,50,0.3 0.2 0.3 0.2,5,"GAAGG","130",3708500
1000,50,0.3 0.2 0.3 0.2,6,"AGAAAA","45",4796600
1000,100,0.25 0.25 0.25 0.25,4,"GATG","439",8191700
1000,100,0.25 0.25 0.25 0.25,5,"CGTCA","125",4737200
1000,100,0.25 0.25 0.25 0.25,6,"TGACTA","40",5848200
1000,100,0.3 0.2 0.3 0.2,4,"GGGG","843",4367600
1000,100,0.3 0.2 0.3 0.2,5,"AGAAA","281",6400400
1000,100,0.3 0.2 0.3 0.2,6,"GGGAAA","92",8037000
5000,50,0.25 0.25 0.25 0.25,4,"ATCT","1023",15511200
5000,50,0.25 0.25 0.25 0.25,5,"CAAGT","276",14220700
5000,50,0.25 0.25 0.25 0.25,6,"CCAATC TCTAGA","82 82",13864400
5000,50,0.3 0.2 0.3 0.2,4,"GAAA","1943",10965200
5000,50,0.3 0.2 0.3 0.2,5,"GAAAG","624",14464900
5000,50,0.3 0.2 0.3 0.2,6,"AAAGGA","188",17920700
5000,100,0.25 0.25 0.25 0.25,4,"CTGA","2035",36224900
5000,100,0.25 0.25 0.25 0.25,5,"TCTTC","535",39996500
```

Image 2: Experiment Results 1

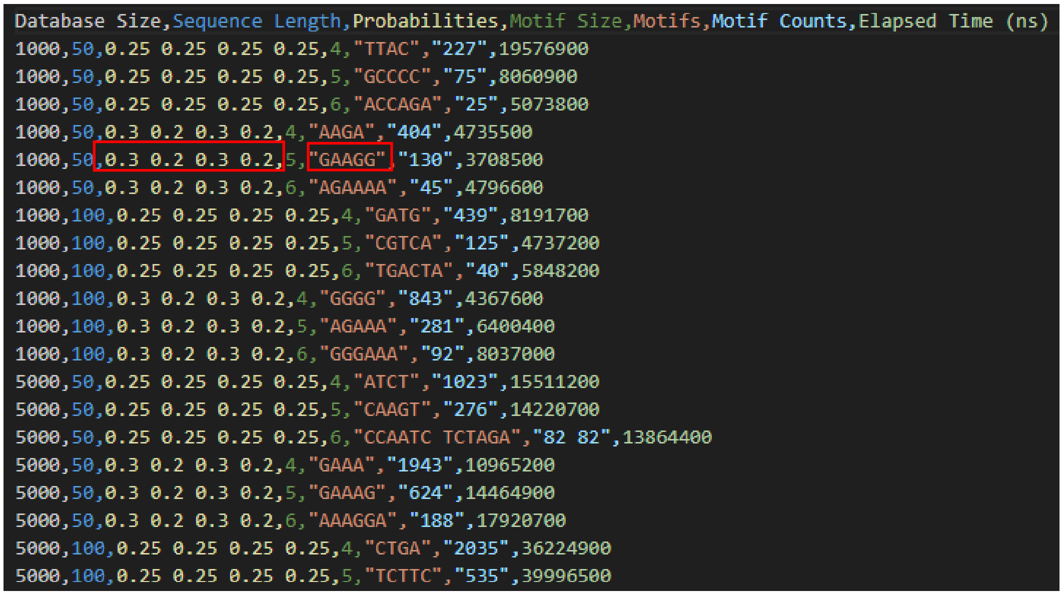
2. **Size of the motifs:** Increasing the motif length tends to reduce the number of frequent motifs since longer motifs are less likely to recur frequently in random sequences.



```
Database Size,Sequence Length,Probabilities,Motif Size,Motifs,Motif Counts,Elapsed Time (ns)
1000,50,0.25 0.25 0.25 0.25,4,"TTAC","227",19576900
1000,50,0.25 0.25 0.25 0.25,5,"GCCCC","75",8060900
1000,50,0.25 0.25 0.25 0.25,6,"ACCAGA","25",5073800
1000,50,0.3 0.2 0.3 0.2,4,"AAGA","404",4735500
1000,50,0.3 0.2 0.3 0.2,5,"GAAGG","130",3708500
1000,50,0.3 0.2 0.3 0.2,6,"AGAAAA","45",4796600
1000,100,0.25 0.25 0.25 0.25,4,"GATG","439",8191700
1000,100,0.25 0.25 0.25 0.25,5,"CGTCA","125",4737200
1000,100,0.25 0.25 0.25 0.25,6,"TGACTA","40",5848200
1000,100,0.3 0.2 0.3 0.2,4,"GGGG","843",4367600
1000,100,0.3 0.2 0.3 0.2,5,"AGAAA","281",6400400
1000,100,0.3 0.2 0.3 0.2,6,"GGGAAA","92",8037000
5000,50,0.25 0.25 0.25 0.25,4,"ATCT","1023",15511200
5000,50,0.25 0.25 0.25 0.25,5,"CAAGT","276",14220700
5000,50,0.25 0.25 0.25 0.25,6,"CCAATC TCTAGA","82 82",13864400
5000,50,0.3 0.2 0.3 0.2,4,"GAAA","1943",10965200
5000,50,0.3 0.2 0.3 0.2,5,"GAAAG","624",14464900
5000,50,0.3 0.2 0.3 0.2,6,"AAAGGA","188",17920700
5000,100,0.25 0.25 0.25 0.25,4,"CTGA","2035",36224900
5000,100,0.25 0.25 0.25 0.25,5,"TCTTC","535",39996500
```

Image 3: Experiment Results 2

3. **Varying probabilities:** Modifying the probabilities for the generation of each base affects the motif distribution. For instance, with equal probabilities for each base motif are more evenly distributed.



```
Database Size,Sequence Length,Probabilities,Motif Size,Motifs,Motif Counts,Elapsed Time (ns)
1000,50,0.25 0.25 0.25 0.25,4,"TTAC","227",19576900
1000,50,0.25 0.25 0.25 0.25,5,"GCCCC","75",8060900
1000,50,0.25 0.25 0.25 0.25,6,"ACCAGA","25",5073800
1000,50,0.3 0.2 0.3 0.2,4,"AAGA","404",4735500
1000,50,0.3 0.2 0.3 0.2,5,"GAAGG","130",3708500
1000,50,0.3 0.2 0.3 0.2,6,"AGAAAA","45",4796600
1000,100,0.25 0.25 0.25 0.25,4,"GATG","439",8191700
1000,100,0.25 0.25 0.25 0.25,5,"CGTCA","125",4737200
1000,100,0.25 0.25 0.25 0.25,6,"TGACTA","40",5848200
1000,100,0.3 0.2 0.3 0.2,4,"GGGG","843",4367600
1000,100,0.3 0.2 0.3 0.2,5,"AGAAA","281",6400400
1000,100,0.3 0.2 0.3 0.2,6,"GGGAAA","92",8037000
5000,50,0.25 0.25 0.25 0.25,4,"ATCT","1023",15511200
5000,50,0.25 0.25 0.25 0.25,5,"CAAGT","276",14220700
5000,50,0.25 0.25 0.25 0.25,6,"CCAATC TCTAGA","82 82",13864400
5000,50,0.3 0.2 0.3 0.2,4,"GAAA","1943",10965200
5000,50,0.3 0.2 0.3 0.2,5,"GAAAG","624",14464900
5000,50,0.3 0.2 0.3 0.2,6,"AAAGGA","188",17920700
5000,100,0.25 0.25 0.25 0.25,4,"CTGA","2035",36224900
5000,100,0.25 0.25 0.25 0.25,5,"TCTTC","535",39996500
```

Image 4: Experiment Results 3

DISCUSSION OF RESULTS

The experiments demonstrated how parameter variations affect both computational performance and motif detection. The time complexity grows linearly with the number of sequences and sequence length. This is clearly if we check the increase in runtime as database grows. Motif detection becomes more computationally expensive with larger datasets.

In addition, the choice of motif length also plays a crucial role in determining the results. Short motifs are more likely to appear frequently, while longer motifs are less common.

The variability in results leads in the expected behavior in stochastic systems and adds a layer of realism, as biological systems are inherently non-deterministic.

CONCLUSIONS

In summary, the project successfully implements a motif search algorithm for DNA sequences, whose complexity is directly linked to the input parameters, in particular the number of sequences and their length. The systemic approach of the project allows the study of DNA sequence patterns under different conditions, providing a flexible framework for experimentation.

Finally, reviewing the project from the perspective of chaos theory, due to the variability of the introduced parameters uncertainty is generated, which could translate into chaos to a greater or lesser extent.

However, this chaotic element reflects the real complexity of biological systems, in which randomness and unpredictability are inherent characteristics. The system effectively models such behavior, providing valuable insights into how different parameters affect both motif detection and computational performance. In future iterations, refining the balance between randomness and control could lead to more optimized and predictive results, improving the system's practical applications in biological research.