UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Systems Analysis Workshop 1 report

Engineering faculty

Systems Engineer

University Francisco Jose de Caldas

Andrey Camilo Gonzalez Caceres

20231020070

Bogotá DC

2024

Introduction

This report contains information about workshop 1, which deals with programming a code that creates an artificial database in a .txt file that generates according to the parameters of the length of the string with a range from 5 to 100 characters presented per sequence and how many times said string will be generated, saving certain parameters from 1000 sequences to 2000000 sequences and another that evaluates according to a chosen candidate the number of times it is presented in the artificial database (previously generated .txt file). The analysis for this system will also be shown, presenting the systemic analysis, complexity analysis, chaos analysis, results, discussion of results and conclusions.

Systemic analysis

System description:

This system consists of some codes in the Java language, one of which generates genetic sequences according to the established parameters, the other determines the appearance of a certain established and modifiable motif and how long this process takes, all this by reading a .txt file taking it as if it were an artificial database of genetic sequences.

Elements:

- Motif.
- Genetic sequences.
- Motif occurrences.
- Artificial database file.

Relations:

- The Artificial database file contains the genetic sequences.
- The genetic sequences could contains motifs.
- The motif occurrences counts the motifs present in the artificial database file.
- The Artificial database file contains a lot of Motifs.

Complexity analysis

This system can be so complex and it can´t be, it depends on the variable values that we define in the code, for example, if we choose an A big value like 0.91, and an C,T or G value like 0.03, the A letter will be the letter that is going to appear the most, for example "AAAACAAATAAAGAAAAAAAA" and it isn´t a complex system, but if each letter has the value (0.25) we can observe that every genetic sequence will be so variable, for example, "ACGTTCGATCGATAGCT" and it makes the system more complex.

The process time increases or decreases according to the values set for the variables.

Chaos system

We can see that the system can be quite predictable according to the results, since the program that creates the genetic sequences and evaluates the number of motifs presented is a very similar value according to the parameters established for the creation and reading of the artificial database of genetic sequences.

Results

Different artificial datasets:

Database size	Base probabilities (A,C,G,T)	Motif size	Motif	Motif occurrences	Time to find Motif
100000*10	(0.5, 0.25, 0.15, 0.1)	5	ACTGA	535	32 ms
200000*15	(0.2, 0.2, 0.2, 0.4)	6	ACTGAC	247	52 ms
500000*30	(0.25. 0.25, 0.25, 0.25)	5	ACTGA	12487	94 ms
1000000*45	(0.3, 0.3, 0.2, 0.2)	4	ACTG	151523	172 ms
1500000*80	(0.15, 0.45, 0.18, 0.22)	7	ACTGACT	4361	285 ms
2000000*100	(0.25, 0.25, 0.25, 0.25)	4	ACTG	758521	545 ms

Conclusions

In this project, I generated artificial genetic sequences and used algorithms to detect motifs in the data. I managed to optimize the pattern search by measuring the execution time and applying techniques to improve efficiency in large databases.

In addition, I performed experiments varying the database size, base probabilities, and motif size, and summarized the results in a table that includes the number of occurrences and the time to find the motif observed in the document. This work allowed me to learn about bioinformatics, algorithm optimization, and data analysis.