SYSTEMS ANALYSIS GROUP 020-84

REPORT WORKSHOP 1

TEACHER: CARLOS ANDRÉS SIERRA VIRGUEZ

LAURA SOFÍA CULMA OSPINA (20231020163)

UNIVERSIDAD DISTRITAL FRANCISO JOSÉ DE CALDAS

SYSTEMS ENGINEER

2024

SYSTEMIC ANALYSIS

The main purpose of this project is analyzing the behavior of motifs in the four fundamental elements in the DNA (adenine (A), thymine (T), guanine (G) and cytosine (C)). For do that we can separate the whole project in these parts:

1. Elements:

The elements in the project are the classes and the archive, in my case there are 5 classes:

- EntropyFilter
- FileReaderUtil
- Launcher
- Motif
- MotifFinder
- Database.txt

2. Relations (General):

1) EntropyFilter: Do a filter of all the sequences before found the motif
2) FileReaderUtil: Provides a methos who reads the sequences generated
3) Launcher: is the “controller” that call all the classes.
4) Motif: generates the data (sequences) and then are processed by FileReaderUtil and MotifFinder
5) MotifFinder: Performs the main analysis of searching motif in the sequences
6) database.txt: Save the aleatory sequence of DNA

The next graph represents the relations between the elements:

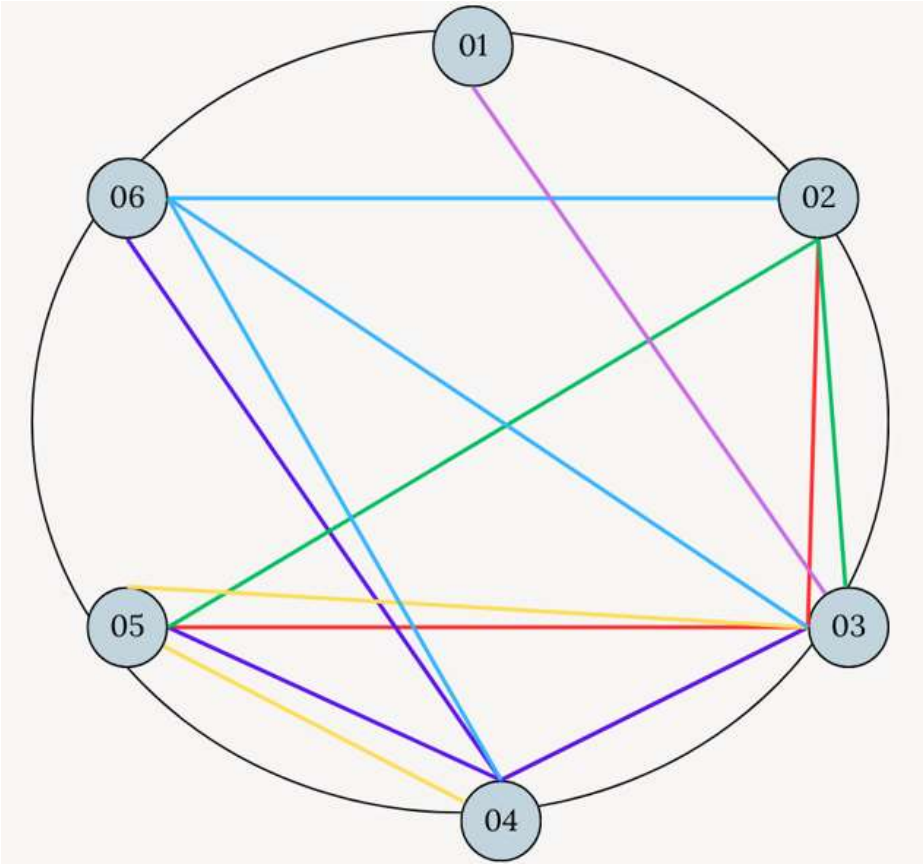

Image 1. Relations between elements

Red: launcher

- 3-4: Calls a motif to generate a database of aleatory sequences of DNA
- 3-2: Use FileReaderUtil to read the sequences of DNA
- 3-5: Invoke MotifFinider to find the most concurrence motif

Dark purple: Motif

- 4-6: Generate then sequences of DNA according the probabilities of the nitrogen bases
- 4-3: Is used by launcher to create a database of sequences
- 4-5: Generate the sequences for be analyzed for MotifFinder to search commons motifs

Dark green: File Reader

- 2-3: Is used by launcher to upload the sequences generated by Motif from a file
- 2-5: Provides the sequences to MotifFinder

Yellow: MotifFinder

- 5-3: It is used by Launcher to identify the most frequent motif in the sequences provided by FileReaderUtil

- 5-4: The set of sequences that Motif generated is analyzed, looking for the most repeated motif.

Light blue: database

- 6-4: The Motif class generates the database.txt file using the generateDatabase method
- 6-2: The FileReaderUtil class is responsible for reading the contents of the database.txt file using the readDatabase method
- 6-3: The Launcher class invokes the entire process, starting with the generation of the database.txt file

Purple - pink: entropy

- 1-3 Is executed by launcher

3. Inputs:

- The size of the sequences of the artificial database (int)
- The size of each sequence (int)
- The size of the motif (int)
- The weights of each base (the probability for adenine, thymine, guanine and cytosine) (double)

4. Process:

- Create an artificial database for the DNA sequences. In this case is a txt archive.
- Calculate the entropy for do a filter of the sequences that don't achieve the criteria. I do that with a class who calculates how diverse are the sequences.
- Obtain the motif of every case. Using a class with an important variable (maxMotif) we find the motif that has most occurrences.
- Count the motif occurrences. As we did find the motif, I create a variable who save the count.
- Calculate the time to find the motif: I create two variables, one for the start of the time, and other for the end of the time.

5. Output:

- The motif (String)
- The occurrences of the motif (int)
- The time to find the motif (long)

COMPLEXITY ANALYSIS

The complexity of the analysis of this system depends of the inputs, because all can change because of that. The complexity depends directly of the amount of the sequences, the length of each sequence and the size of the database.

The data flow is sequential, which simplifies control, but does not optimize the total processing time, since the stages cannot be executed in parallel. The total execution time depends on the sum of the pattern generation, reading, processing and analysis operations.

CHAOS ANALYSIS

As we said before, small changes can alter the expected results, tiny variations in the weights of the DNA sequences can provoke different results and motif totally different. In this part we can identify a butterfly effect because all in the process would work well until the end of the procedure

Is a system really sensible, anything can change the outputs. Just with looking one time the database we see a lot of letters with no sense, we see disorder, but our system takes that disorder and transform in a sequence of patterns and take the most repetitive one. If our database were only with one or two nitrogen bases, probably our system would be more organized, but it wouldn't make sense.

Even though the inputs are established, the results are not, because all the sequences are formed by a random object. And that makes our system not deterministic, the motif, the time, the occurrences are unpredictable, we can make a possibility and a statistic but nothing is sure.

Shannon's formula for entropy is a fundamental measure in information theory that quantifies the level of uncertainty or disorder in a data set. And there is an emergent relationship between disorder (entropy) and regularity (motifs). In sequences with greater disorder, the probability of finding repetitive motifs may decrease, while in more ordered sequences, the appearance of repetitive patterns increases. Exists a balance between the disorder and order, and that is what chaos theory poses.

RESULTS

The next table shows the results of changing the weights in the nitrogen bases.

Database Size	A	C	G	T	Motif Size	Motif	Motif Ocurrences	Time to Find Motif
100000	0,25	0,25	0,25	0,25	6	TTTCAC	1222	5,0566
100000	0,25	0,25	0,25	0,25	6	CTTATG	1217	3,5502
100000	0,25	0,25	0,25	0,25	6	GTACGG	1225	3,3806
100000	0,25	0,25	0,25	0,25	6	GTTTTA	1229	5,6001
100000	0,26	0,20	0,27	0,26	6	GTGTGT	1864	5,4369
100000	0,27	0,20	0,27	0,26	6	AGGGAG	1821	3,5301
100000	0,23	0,23	0,31	0,23	6	GGGGGG	3965	5,3696
100000	0,28	0,24	0,24	0,24	6	AAAAAA	2292	5,0518
100000	0,26	0,26	0,22	0,26	6	ATCTTC	1518	5,7488
100000	0,27	0,19	0,27	0,27	6	GATTAG	1885	3,4939
100000	0,34	0,22	0,22	0,22	6	AAAAAA	7028	4,0169

Table 1. Changes in the weights od the nitrogen bases

The next table shows the results of changing the size of the nitrogen bases.

Database Size	Probability of Bases	Motif Size	Motif	Motif Ocurrences	Time to Find Motif
100000	0,25	6	CGAGCG	1224	3,1675
80000	0,25	6	GGGAGT	969	3,1223
70000	0,25	6	CTAGGT	868	3,5558
60000	0,25	6	ACTACA	750	2,3196
50000	0,25	6	CGCATT	635	1,9619
100500	0,25	6	TATCCC	1206	3,4156
101000	0,25	6	CGCGTC	1227	3,7455
300000	0,25	6	AAAGTA	3520	9,2372
400000	0,25	6	TTTAGA	4642	13,5948
450000	0,25	6	TCCTCA	5215	12,8163

Table 2. Changes in the size of the database

The next table shows the results of changing the size of the motif.

Database Size	Probability of Bases	Motif Size	Motif	Motif Ocurrences	Time to Find Motif
100000	0,25	4	GGGG	83863	13,9588
100000	0,25	4	TATC	83343	12,5013
100000	0,25	5	TGGCA	20686	12,9796
100000	0,25	6	TTGGAC	5181	13,1086
100000	0,25	7	ACGCCCA	1351	21,5472
100000	0,25	8	GGAATCAA	380	25,0454
100000	0,25	9	CTTGAACGC	112	25,5898
100000	0,25	9	ATGGCTATT	116	23,404
100000	0,25	10	ACTAGACCGT	42	24,2886
100000	0,25	10	TCGCGGTACT	41	27,0478

Table 3. Changes in the motif size

DISCUSSION OF RESULTS

1. **For the first table:** As we expected when we increase the weights in the bases, the motif has with more repetition that letter. If two of them or three of them have the same value and is the biggest, the three could appear in the motif (for example the nine result). The interesting part is when all the bases have the same weigh, is curious that in the four results, is always a base who has predominancy and in the in the first three repeats three times It appears that it is quite rare for any base to repeat more than three times in such cases.

2. **For the second table:** For the second table, we observe that the motif occurrences increase in direct proportion to the size of the database. While we might expect that the time required to find the motifs would scale similarly, this is not always the case. The processing time depends on various factors, such as the computer's performance and the number of sequences identified. Consequently, time is not a deterministic value, and fluctuations in computing power or database structure can lead to variations in the time required for motif discovery.

3. **For the three table:** The relationship between motif size and motif occurrences is inversely proportional because a larger "pattern" has fewer chances of appearing frequently. As the motif size increases, it becomes harder to find many matches. Over time, this effect becomes even more pronounced, as the search process takes longer and the number of matches decreases. With the time this time varied a lot, I would say that when the motif size is larger, is more difficult to find more like them.

CONCLUSIONS

1. As base weights increase, these bases appear more frequently in the motif. When the weights are equal, one base usually dominates and is repeated three times. We can say that exists certain dominance in the system
2. The occurrences increase with database size. However, processing time varies due to factors such as computer performance, indicating that the time is not deterministic and can fluctuate.
3. There is an inverse relationship between the size of the motif and the number of occurrences: larger motifs are less frequent.
4. Small variations in input parameters can result in large differences in results, unleashing the butterfly effect. The system's sensitivity to these changes demonstrates that complex patterns can emerge from chaotic data.
5. More disorder in sequences makes it less likely to find repeating motifs, while more order leads to more regular patterns. Shannon's theory helps measure and manage this disorder, allowing us to better organize and understand the information.