Technical report Workshop 1 System Analisis

Angel Andres Diaz Vergara - *aadiazv@udistrital.edu.co*

1. INTRODUCTION

For this workshop we have the task of generating and analyzing artificial biological datasets, consisting of DNA sequences. It focuses on the construction of an artificial database composed of nucleotide sequences (A, C, G, T) with customizable probabilities for each base, simulating the random nature of biological data. The objectives are several: first, to efficiently generate these sequences using the available programming knowledge. Second, to identify recurrent motifs of different sizes, which are short patterns of interest in genomics. Third, make a chaotic system, using Shannon entropy to filter excessively repetitive sequences, to increase randomness and diversity in the dataset. Fourth, through experimentation, the results and system behavior are compared with different variables, both before and after entropy-based filtering.

2. Systemic Analysis

As the workshop revolves around generating a simulated biological dataset composed of nucleotide sequences, performing motif search operations, and entropy filtering as basic principles, the system can be divided into three main components:

2.1. Artificial database generation

The objective is to create an artificial database containing between 1,000 and 2,000,000 sequences of lengths between 5 and 100, each sequence is composed of the four nucleotide bases (A, C, G, T) with probabilities defined as parameters, allowing variation in the frequency of nucleotide occurrence, also the generated database must be stored in a .txt file, suggesting that the system must perform file writing operations and variable customization for different probabilities, in addition a distributed computing or divide and conquer strategy is suggested to handle the generation and storage of this large amount of data, indicating the need for the system of scalability and efficiency.

2.2. Motif search algorithm

It focuses on iterating over the generated sequences to find motifs of a given size (between 4 and 10 nucleotides), so the system must iterate through all combinations of nucleotide bases of the given size and identify the most frequent motif, and given the large amount of data it also suggests that distributed computing and efficient search methods should be used.

2.3. Entropy filtering

The objective is create a chaotic system, so we seek to avoid repetitiveness and promote diversity in the sequences, for that, we must measure and control the level of randomness (chaos) in the sequences, for this it is suggested to use Shannon entropy as a filtering method, where highly repetitive sequences are filtered to improve the diversity of the remaining data set.

3. Complexity Analysis

The complexity of this system is multifaceted and ranges from the computational complexity in the generation of sequences to the algorithmic complexity in the search for motifs and the calculation of entropy.

3.1. Sequence generation

The generation of sequences has a very high computational complexity, since the random factor is often needed, in addition to the fact that the configurable probabilities of the nucleotides add more complexity and control of character strings is needed to give the result of the final sequence.

3.2. Motif search

The motif finding algorithm requires generating all possible combinations of nucleotide bases of size s, which has a complexity based on s, which can be very high, and for each motif, a comparison must be made with all the sequences in the database, resulting in very high complexity.

3.3. Entropy calculate

Shannon entropy can be calculated for each sequence based on the frequency of nucleotide bases and this involves calculating the probability distribution of nucleotide bases and applying the entropy formula: $H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$ so, this result in very high complexity

4. Chaos Analysis

In this context, chaos refers to the diversity and unpredictability of the generated nucleotide sequences, the system is desired to be chaotic because the more diverse sequences are the better to analyze, but the level of randomness in a computer is very low, wich causes that the first generated dataset have low diversity and randomness, for this porpuse shannon entropy is use, to mesure the randomness fo each secuence, defined by the formula: $H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$ where $p(x_i)$ represent the probability of the nucleotide in the secuence, high entropy indicates a higher degree of randomness and diversity in the sequence, so for the low randomness of the initial generated dataset, there a filter with this entropy to save the secuences with higher entropy for better analisis.

5. Results

Below are the results obtained for different variables

Database Size	Probability of Bases	Motif Size	Motif	Motif Occurrences	Time to Find Motif
158315	A=0.25, C=0.25, G=0.25, T=0.25	4	TACT	30429	234 ms
1622578	A=0.25, C=0.25, G=0.25, T=0.25	5	TCTGG	76711	2204 ms
374480	A=0.25, C=0.25, G=0.25, T=0.25	6	ACCTTG	4518	570 ms
466537	A=0.25, C=0.25, G=0.25, T=0.25	7	GCATTTT	1458	749 ms
701831	A=0.25, C=0.25, G=0.25, T=0.25	8	AACACCTC	579	1860 ms
194220	A=0.25, C=0.25, G=0.25, T=0.25	9	GCGTAAGCA	58	653 ms

Table 1. results with generated dataset

Database Size	Probability of Bases	Motif Size	Motif	Motif Occurrences	Time to Find Motif
158315	A=0.25, C=0.25, G=0.25, T=0.25	4	CTGA	10155	65 ms
1622578	A=0.25, C=0.25, G=0.25, T=0.25	5	CTGAC	46349	1053 ms
374480	A=0.25, C=0.25, G=0.25, T=0.25	6	GCTAGC	3616	359 ms
466537	A=0.25, C=0.25, G=0.25, T=0.25	7	TACGTCA	1276	500 ms
0	A=0.25, C=0.25, G=0.25, T=0.25	8		0	0 ms
194220	A=0.25, C=0.25, G=0.25, T=0.25	9	CGAATGAGC	53	304 ms

Table 2. results with filtered dataset

Database Size	Probability of Bases	Motif Size	Motif	Motif Occurrences	Time to Find Motif
884763	A=0.05, C=0.7, G=0.2, T=0.05	6	CCCCCC	3810646	1099 ms
1028017	A=0.1, C=0.4, G=0.4, T=0.1	7	CCCGCCC	77794	1505 ms
884196	A=0.1, C=0.2, G=0.2, T=0.5	8	TTTTTTTT	153087	1859 ms
1356535	A=0.5, C=0.2, G=0.2, T=0.1	9	AAAAAAAAA	115955	4400 ms

Table 3. results with generate dataset and different probabilities

Database Size	Probability of Bases	Motif Size	Motif	Motif Occurrences	Time to Find Motif
884763	A=0.05, C=0.7, G=0.2, T=0.05	6	CAGTCC	1660	243 ms
1028017	A=0.1, C=0.4, G=0.4, T=0.1	7	GCATCGC	2229	791 ms
884196	A=0.1, C=0.2, G=0.2, T=0.5	8	TTCGATCT	974	669 ms
1356535	A=0.5, C=0.2, G=0.2, T=0.1	9	AAATCGAAA	798	1306 ms

Table 4. results with filtered dataset and different probabilities

6. Discussion of Results

The results show variability in the execution time and the number of motifs occurrences, depending on the size of the data set and the probability of each nucleotide, the most observed patterns are:

-In the case of nucleotides with equal probabilities, the most repeated motifs are randomly composed, although several repeated elements are seen, with nucleotides that have a higher probability of appearance being composed of that nucleotide.

-The time required to find the motifs tends to increase with the size of the motif and the dataset.

-Different probability values for nucleotides affect both the number of repeats and the search time.

7. Conclusions

The tests performed show that data set size, motif size, and nucleotide probabilities are key factors affecting both the number of motif occurrences and the search time required, the main conclusions are:

-When the nucleotide probability is the same, certain patterns of the same nucleotides tend to be repeated at time to generate a sequence, indicating that the randomness with which it is generated is low

-The probability of each nucleotide has a significant impact on the repetition of motifs. An increase in the probability of a specific nucleotide usually results in a higher number of motifs occurrences containing that nucleotide.

-The time to find motifs increases with the size of the motif and the data set, which is consistent with the complexity of searching for patterns in large volumes of data.

-Unbalanced probability distributions can lead to a concentration of repeated motifs which can affect the interpretation of the data.