Detection of motives in a dataset

Juan David Campos Salcedo

Faculty of Engineering

Universidad Distrital Francisco Jose de Caldas

Bogotá, Colombia

jdcamposs@udistrital.edu.co

Detection of motives in a dataset

1. **Abstract:** *This report seeks to use different sets of artificial data and different generation probabilities for each base and different amounts of sequences to look for motives in the generated sequences and analyze the data to see how the behavior can vary if a Shannon filter is applied and how this can affect the performance of the system.*

2. ***Systemic analysis:*** The implemented code is used to search for motifs in the DNA chain generated by it; to do this, the tasks are distributed to achieve this goal. These are carried out within the code in different sections to achieve greater agility during the generation of these DNA chains and to search for motifs within the same generated sequences.

   The program is divided into 2 classes, the "main" class by which the order that the procedures must follow will be given, and the bioinformatics class, which saves and executes the procedures that the system must do to achieve its objective of finding the motives, these procedures within this class are:

   - ***Dataset creation:*** through some values entered into the system, the sequences of the dataset are determined, and what the dataset should have is structured.
   - ***Obtaining the motif:*** from a search and generation of combinations, a search is made for the motif, how many motifs there are and the time it took to obtain that data.
   - ***Save data:*** data is saved to process it and perform a more exhaustive analysis of it.
   - ***Application of entropy:*** To ensure that there is more variety between strings and that they do not have two characters in a row, Shannon's entropy is applied to the sequences in the dataset and then a new dataset is created with the above to have a point of comparison between the two and to analyze how they differ after this process has been carried out.

3. ***Complexity analysis:*** Within what is possible and how the system is covered, the complexity lies in the search for the motives and the creation of the dataset itself. It is necessary to ensure at the time of creation of the dataset that it does not go beyond the established parameters through exceptions that prevent the code from collapsing due to compilation errors or data reception errors. These details must be kept in mind so that the program can continue its course without problems and so

that the system does not, due to some defect, lead to a series of events taking place, which could negatively affect the internal structure of the system. Thus, it will continue its course and can achieve its final purpose.

It is also worth emphasizing that the complexity will lie in the values entered by the user, since these will generate more or less chains of sentences, which will depend on the number of motives found within these sentences and the speed of operation of how long it will take to find these same sentences.

4. ***Chaos analysis:*** The level of chaos of the sequences generated within the dataset will be influenced by two factors, that the code works correctly and that the strings are filtered with Shannon's entropy, on the one hand, the program must be well structured and ensure that it is fail-safe, because if a failure is found in the system, it causes a domino effect that causes failures in the system, so we must ensure that this does not happen by correctly restricting all possible cases that could generate an effect like this, while Shannon's entropy from a formula which encourages that if any character is repeated too much within the same sequence consecutively there is more variety and unforeseen values that encourage chaos.

5. ***Results***: From the code, 10 databases will be generated, with different values for the size of the sequences and the chains, which are shown below:

Database Size	Motif Size	Motif	Motif Ocurrences	Time to Find Motif. (seconds)
2345	7	GAGTTAA	30	0.036 s
23884	10	CATCACCGCG	12	0.754 s
46735	8	CTGGTGAT	100	0.491 s
64737	4	TTTG	25525	0.215 s
85473	5	ACGTC	8352	0.345 s
232435	6	CTACGC	5661	1.003 s
248734	9	CAGCCATTA	129	5.080 s
284365	9	CAAAGTTGG	141	5.739 s
388453	7	CGCTGAT	2316	2.635 s
938473	4	CTGT	364264	3.367 s

and if the entropy filter is applied to the same database, the results are as follows:

Database Size	Motif Size	Motif	Motif Occurrences	Time to Find Motif. (seconds)
2345	7	GAGTTAA	30	0.070 s
23884	10	CATCACCGCG	12	2.038 s
46735	8	CTGGTGAT	100	1.517 s
64737	4	TTTG	25469	0.634 s
85473	5	ACGTC	8364	1.007 s
232435	6	CTACGC	5666	3.204 s
248734	9	CAGCCATTA	130	15.114 s
284365	9	CAAAGTTGG	142	17.313 s
388453	7	CTACTGA	2319	8.180 s
938473	4	ACTG	364693	3.367 s

6. ***Discussion of results:*** What can be observed in the results is that with the application of entropy, the values tend to increase or decrease depending on the size of the motif sequence and the size of the dataset, and that in some cases with the variations of the characters after the filter some motifs also changed, meaning that where there were extreme changes in the motif, the generated sequences had a low level of chaos. What is also observed is that with the implementation of the filter, it produced a considerable increase in the effective time with which the motifs are searched.

7. ***Conclusions:*** With the above mentioned, it can be stated that the higher the values entered at the beginning to form the dataset, the chaos will be present to a greater extent, and consequently when passing the dataset through a Shannon entropy filter, it will not have significant changes at the time of the levels of chaos that can be presented within these and consequently it will not have major changes with respect to the original results, except that it will take some time and its performance will decrease when searching for these same results, since the search has a greater complexity, while on the contrary, between lower values, it will have less initial chaos and at the time of passing through the filter it has more possibilities of having value changes and a greater possibility of increasing its performance time.