Entropy and Motif Analysis in Genetic Sequences Using Divide and Conquer

Laura Valentina Cubillos Acero
Universidad Distrital Francisco José de Caldas
Ingeniería de Sistemas
Email: lvcubillosa@udistrital.edu.co

***Abstract*—This report presents the development of a system that generates artificial genetic sequences and finds the most frequent motifs (patterns) within them. Additionally, Shannon entropy is used as a chaos measure to filter repetitive sequences and improve motif analysis. The approach followed is a "divide and conquer" algorithm to optimize tasks.**

I. Introduction

The analysis of genetic sequences is a key area in bioinformatics. Detecting repeated patterns or motifs in DNA sequences can reveal important biological characteristics. In this work, a system is developed that:

- Generates a database of artificial genetic sequences.
- Finds frequent motifs of a given size.
- Calculates Shannon entropy to filter repetitive sequences.

The system is modular and based on divide and conquer techniques, optimizing both the generation and analysis of data.

II. Systemic Analysis

The developed system generates random genetic sequences based on the bases 'A', 'C', 'G', 'T' with adjustable probabilities. Subsequently, the system detects the most frequent motif in these sequences, and finally, applies Shannon entropy to filter sequences with low diversity.

The system structure includes the following steps:

1) Generation of sequences with specific probabilities for each base.
2) Search for most frequent motifs of a fixed size.
3) Calculation of Shannon entropy to measure sequence chaos.

III. Complexity Analysis

A. Sequence Generation

The generation of sequences has a complexity of $O(n \cdot m)$, where n is the number of sequences and m is the length of each sequence.

B. Motif Search

The motif analysis involves searching for k-length substrings in all sequences. This process has a complexity of $O(n \cdot m \cdot k)$.

C. Entropy Calculation

The calculation of entropy for each sequence is linear with respect to the sequence length, resulting in a complexity of $O(n \cdot m)$.

IV. Chaos Analysis (Entropy)

Shannon entropy is used to measure randomness in the sequences. The higher the entropy, the more diverse the sequence. We used a threshold of 1.5 to filter sequences with low entropy, which tend to have many repetitions of the same base.

V. Results

Several experiments were conducted with different base probabilities and motif sizes. The following tables summarize some of the obtained results.

TABLE I: Results without entropy filtering

DB Size	Probabilities	Motif Size	Motif	Occurrences	Time (ms)
1000	A: 0.25, C: 0.25, G: 0.25, T: 0.25	5	ACGTA	15	120
2000	A: 0.25, C: 0.25, G: 0.25, T: 0.25	5	TGACA	12	200
1000	A: 0.40, C: 0.20, G: 0.20, T: 0.20	5	AAAAC	22	150

After applying the entropy filter, the results were as follows:

TABLE II: Results with entropy filtering

Filtered DB Size	Entropy Threshold	Motif Size	Motif	Occurrences	Time (ms)
800	1.5	5	CGTAC	9	110
1500	1.5	5	GATCA	10	180

VI. Discussion of Results

By changing the probabilities of the bases, the most frequent motif varied significantly. Additionally, after applying the entropy filter, the dataset size was reduced, allowing the discovery of more chaotic motifs and reducing execution time.

VII. Conclusions

The divide and conquer approach is suitable for genetic sequence analysis problems, as it allows the optimization of different phases of data processing. The use of Shannon entropy as a chaos measure proved effective in eliminating repetitive sequences and improving motif analysis. The overall system performance is acceptable but could be improved for larger datasets using parallelization techniques.

References

1) Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27, 379-423.
2) Mount, D. (2004). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.