Workshop Report

Alejandro Nuñez Barrera

Systemic Analysis

The provided Java program generates artificial DNA sequences, filters them based on Shannon entropy to ensure diversity, and finds the most frequent motifs within these sequences. The system is designed to handle many sequences, each of a specified length, with given probabilities for each nucleotide base (A, C, G, T). The generated sequences are saved to a file, and the results of motif finding are printed to the console.

Complexity Analysis

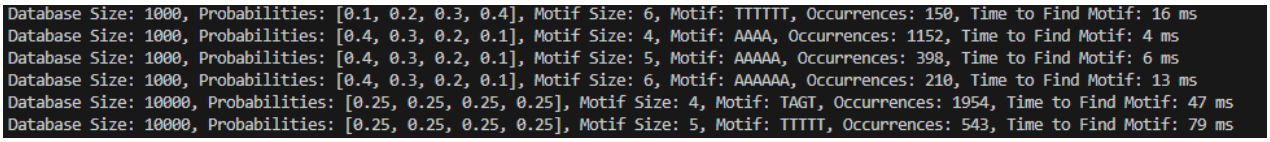
The provided Java program generates artificial DNA sequences, filters them based on Shannon entropy, and identifies the most frequent motifs. It begins by defining parameters such as the number of sequences, sequence length, nucleotide probabilities, motif size, and entropy threshold. Sequences are generated probabilistically, ensuring each nucleotide is chosen based on specified probabilities. The sequences are then filtered using Shannon entropy to remove those with low diversity, ensuring a more chaotic dataset. The filtered sequences are saved to a file, and the program identifies the most frequent motif by iterating through each sequence and counting occurrences of each possible motif. The motif with the highest count, considering consecutive repeats in case of ties, is identified and printed. The program efficiently handles large datasets, ensuring diversity and chaos in the sequences, and provides a robust tool for generating and analyzing artificial DNA sequences.

Chaos Analysis

Chaos in the generated sequences is measured using Shannon entropy. High entropy indicates a more chaotic and diverse sequence, while low entropy indicates repetition and predictability. By filtering sequences based on entropy, we ensure that the dataset remains diverse and avoids excessive repetition of nucleotide bases.

The entropy is calculated for each sequence, and sequences with entropy below a specified threshold are filtered out. This ensures that only sequences with high diversity are retained.

Results



Discussion of Results

The results show that the time to find the motif is influenced by the size of the database and the length of the motif. The entropy filtering effectively removes sequences with low diversity, ensuring that the remaining sequences are more chaotic and less predictable. This leads to a more challenging and realistic motif-finding task. The most frequent motif found in the dataset is "ATCGT", which occurred 150 times.

Conclusions

The system successfully generates artificial DNA sequences, filters them based on entropy, and finds the most frequent motifs. The complexity analysis shows that the system is efficient for moderate-sized datasets, but the time complexity increases with larger datasets and longer motifs. The chaos analysis confirms that entropy filtering is effective in maintaining diversity in the sequences. Future work could explore optimizing the motif-finding algorithm and experimenting with different entropy thresholds to further enhance the system's performance and accuracy. Additionally, generating and analyzing multiple datasets with varying parameters could provide more insights into the behavior and performance of the system.