DNA Compression

Roger Pujol
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
roger.pujol.torramorell@est.fib.upc.edu

***Abstract*—In this project I am going to analyse the algorithm described in the paper "DNA Compression Using Hash Based Data Structure" [1].**

The code of this project is Open Source and can be found in: https://github.com/rogerpt32/dna_compression [2].

I. Summary

The summary of the paper(s), data structure(s) (10%) Nowadays, biologists are producing huge volumes of DNA sequences that makes the genome sequence database grow exponentially. Since the amount of data is getting so big, there is the need to find efficient algorithms to compress it. The regular text compression algorithms doesn't perform very well with DNA sequences. Most of the current DNA compression algorithms exploit the repetitive nature of the DNA sequences, but the algorithm presented

II. Opinion

Your personal evaluation on why the paper(s), data structure(s) is (are) relevant (10%)

III. Experiments

A. Design

The design of a set of experiments to validate the main aspects of the data structure(s) (or the proof of the main theorems in case of a theoretic paper, demo) (15%)

B. Analysis

A critical analysis of the results of the experiments (theoretical results) (15%)

IV. Conclusions

Your personal conclusions (10%)

References

[1] Ateet Mehta and Bankim Patel. DNA compression using hash based data structure. 01 2010.

[2] Roger Pujol. DNA compression. https://github.com/rogerpt32/dna_compression, 2019.