Bioinformatics at ENS This isn't an article

Project Report

A contribution to the Minimal absent words identification problem

Victor Jourdan and Armand Ledoux

After the discovery of CRISPR, one of the most important discoveries in the field of bioinformatics, working as an adaptive immunity mechanism, used by bacteria to fend off phages and predatory plasmids. To avoid being eliminated by this biological antivirus, plasmids and viruses have developed strategies, such as avoiding distinct patterns, and developed an internal database, like CRISPR. We tried to design an algorithm to efficiently identify minimal absent ords in a plasmid dataset to better understand the patterns potentially targeted by CRISPR.

1 Introduction

The discovery of the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats), a mechanism used by bacteria to fight against phages and plasmids has been a major breakthrough in bioinformatics. Plasmids and viruses have developed counter-measures, such as avoiding certain patterns thanks to an intern database. The goal of this project is to develop efficient algorithms to list the minimal absent words of a DNA sequence or a set of DNA sequences.

1.1 Paper overview

In this paper we present different versions of algorithms to solve the MAW problem. First, we will introduce the concepts manipulated in this paper, then we will present the naïve and unsatisfying solution. Then we will study our various improvements of the naïve solution, with a complexity analysis of them. Comparison in terms of computation time and space usage will conclude our report.

2 Framework

We will consider the following alphabet, $\Sigma = \{\text{A, C, G, T, N}\}$, the four nucleotides and the N, signifiying unknown. A DNA sequence is a string over this alphabet $s \in \Sigma$. A kmer is a substring of length k of s. An absent word is a word that is not a substring of s. A minimal absent word (MAW) is an absent word of s such that every proper substring of it is a substring of s. For a given $p \in [0, 1)$, and a set of sequences S, a pMAW x is a word that is an absent word in at least $p \cdot |S|$ sequences of S, and such that every proper substring of x is not a pMAW of S.

Conflicts of Interest

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report.

Remarks

This project has been jointly led by Armand and Victor. The estimated contribution time is of … There has been no exchange of code with other groups, we only had brief theoretical discussions with Erwan about different strategies.

Notes and References

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat.

*Corresponding author