Minimal Relative Absent Words as Candidate Host-Pathogen Signatures for Innovative Diagnosis

Sara Almeida[1,2], João Carneiro[3] , Diogo Pratas[1,2,4,5]

1. IEETA - Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Portugal; 2. DETI - Department of Electronics, Telecommunications and Informatics, University of Aveiro, Portugal; 3. Universidade Católica Portuguesa, CBQF - Centro de Biotecnologia e Química Fina – Laboratório Associado, Escola Superior de Biotecnologia, Rua Diogo Botelho 1327, 4169-005 Porto, Portugal; 4. LASI - Intelligent Systems Associate Laboratory, University of Aveiro, Portugal; 5. DoV - Department of Virology, University of Helsinki, Finland.

Introduction

Minimal Relative Absent Words (mRAWs) are the shortest nucleotide sequences **occurring** in one genome, such as a pathogen, while being **absent** from another, such as the host, making them potential genomic **signatures** (1). Recent work on SARS-CoV-2 showed that shorter mRAWs tend to exhibit higher GC-content than both the human and virus genomes and transcriptomes and identified a persistent mRAW in the Spike gene, a key protein involved in host-cell entry, highlighting the relevance of absent words for pathogen detection (2).

Objectives

- Determine whether the GC-content patterns detected with SARS-CoV-2 extend across **distinct** host-pathogen systems;
- **Compare** mRAW profiles using resistant and susceptible host backgrounds.;
- Identify whether recurrent mRAW positional signals map to **biologically relevant** viral regions.

Methods

A **reproducible alignment-free** pipeline was developed to compute and analyze mRAWs from curated host-pathogen FASTA and multi-FASTA datasets. Based on the AltaiR toolkit (3) for RAWs computation, the workflow follows with:

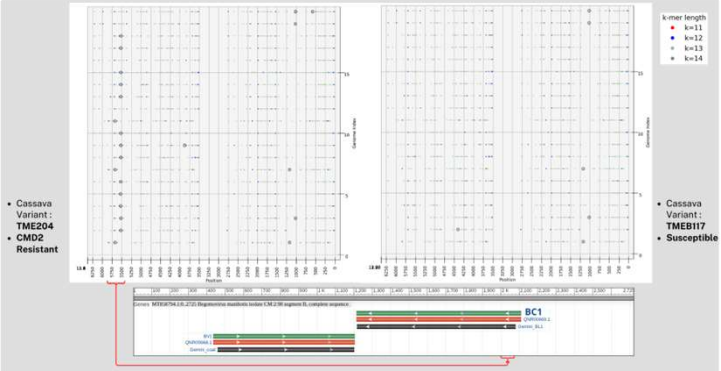
- **mRAWs characterization** according to sequence length, GC-content and genomic position;
- Generation **of standardized visualizations** to support systematic comparisons across pathogen groups and host contexts.

Initial analyses focused on cassava-Cassava Mosaic Virus systems with comparative mRAW profiling analysis between the **CMD2-resistant** cassava genotype **TME204** and the **susceptible** genotype **TMEB117** for one same group of 20 *Begomovirus manihotis* genome sequences.

Results

- A recurrent **12-nt mRAW** with **50% GC-content** was detected at similar genomic positions in **16/20 viral genomes** when **TME204** was used as host background.
- The same recurrent mRAW signal was **not detected** when the susceptible genotype **TMEB117** was used as host background.
- That recurrent region where the mRAW appears mapped to the viral **BC1 gene**, which encodes a **movement protein** involved in the viral spread between plant cells (4).

Top-Down View of 3D mRAWs distribution along the pathogen genome (GC 48-52% marked)



Discussion

Building on previous evidence that mRAWs can reveal informative host-pathogen genomic contrasts, this study explored their behavior in a plant-virus system with direct food-security relevance. The main observation was a recurrent 12-nt mRAW localized within the BC1 region, a gene involved in viral movement. This suggests that mRAW recurrence may provide information beyond sequence composition alone, by highlighting positional patterns in biologically relevant viral regions.

Although no direct relationship with resistance can be inferred at this stage, the detection of this sequence only in the resistant genotype raises space for further investigation within this topic. Future work will test additional cassava genotypes and viral variants, map the sequence onto BC1 protein 3D models and low-complexity profiles, and extend the approach to human-associated host-pathogen systems.

References

1. Silva RM, Pratas D, Castro L, Pinho AJ, Ferreira PJSG. Three minimal sequences found in Ebola virus genomes and absent from human DNA. Bioinformatics. 2015 Aug;31(15):2421–2425.
2. Pratas D, Silva JM. Persistent minimal sequences of SARS-CoV-2. Bioinformatics. 2020 Nov;36(21):5129–5132.
3. Silva JM, Pinho AJ, Pratas D. AltaiR: a C toolkit for alignment-free and temporal analysis of multi-FASTA data. GigaScience. 2024;13:giae086.
4. Aimone CD, Lavington E, Hoyer JS, Deppong DO, Mickelson-Young L, Jacobson A, et al. Population diversity of cassava mosaic begomoviruses increases over the course of serial vegetative propagation. J Gen Virol. 2021;102(7):001622.

This work is carried out within the scope of contract refª 2023.15056.TENURE.043, funded by national funds through the FCT – Foundation for Science and Technology, I.P. This work is also funded by national funds through the FCT – Foundation for Science and Technology, within the scope of UID/50016/2025 and LA/P/0076/2020 (https://doi.org/10.54499/LA/P/0076/2020).