A Multi-Omics and Primer Database for Virus Identification: Focus on HIV, Ebola, and SARS-CoV-2

Lima, AS (1,2); Carneiro, J (3,4); Sousa, SF (5); Sá, VJ (1, 9); Pratas, D. (6,7,8)
(1) LabRP/CIR, School of Health, Polytechnic of Porto, Porto, Portugal; (2) CIIMAR – Interdisciplinary Centre of Marine and Environmental Research of the University of Porto, University of Porto, Matosinhos, Portugal; (3) Centre of Molecular and Environmental Biology (CBMA), Department of Biology, University of Minho, Campus Gualtar, Braga, Portugal; (4) Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Braga, Portugal (5) LAQV/REQUIMTE - BioSIM, Department of Biomedicine, Faculty of Medicine, University of Porto, Porto, Portugal; (6) IEETA/LASI, Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro; (7) DETI, Department of Electronics, Telecommunications and Informatics, University of Aveiro; (8) DV, Department of Virology, University of Helsinki; (9) ALGORITMI/LASI, University of Minho, Braga, Portugal

INTRODUCTION

Highly infectious viruses (e.g.: HIV, Ebola, SARS-CoV-2) pose a threat to global health

- Compile **all variants**;
- Analyse **genomic diversity**;
- Identify **conserved regions**;
- Determine the most reliable **primer pairs** for PCR based on a scoring system.

Better detection methods

AIM

Compile all the data

F.A.I.R DATABASE

Findable, **A**ccessible, **I**nteroperable, and **R**eusable

METHODOLOGY

VIRAL MULTI-OMICS DATA MINING

Get all the complete genomes from NCBI Virus

Upload sequences to Geneious Prime for sequence analysis

- Complete viral genome
- Important documented regions
- Protein originating sequences
- Structural information

Upload data to Galaxy for Multiple Sequence Alignment and Phylogenetic Analysis

MAFFT – Multiple Sequence Alignment
PhyML/RAxML – Phylogenetic tree construction
MEGA/Geneious – Phylogenetic analysis such as cladistics with outgroups

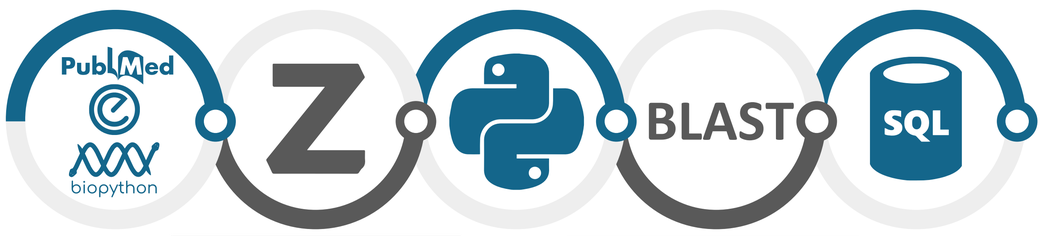
Global conservation scores
Important conserved regions
Evolutive patterns
Etc.

PRIMER RETRIEVAL

Articles are retrieved from **PubMed** through the **Entrez** module

Custom **Python** script for plain text extraction from PDFs and primer scraping

Upload of the validated primers to a **SQL** database



Upload of the file to **Zotero 7** for automatic retrieval of full-text PDFs

Validation of primers scraped against the reference genome through **BLAST**

RESULTS

Retrieved:

- **7304** sequences for **HIV**
- **658** sequences for **Ebola**
- **321916** sequences for **SARS-CoV-2**

01. NotebookLM — **Large Language Model**-based approach. Extraction is done using **user prompts** after uploading a batch of PDFs.

02. DonutAI/OpenLLaMa-7b — **Large Language Model**-based approach. **DonutAI** handles **PDF to plain text** conversion and **OpenLLaMa-7b** handles **primer extraction** from text.

03. RegEx + PyMuPDF4LLM — **Classic** approach. **PyMuPDF4LLM** handles **PDF to plain text** conversion and **primers** are captured using **regular expressions**.

Table 1. Results of the primer scraping process using three different models.

Model	Accuracy (%)	Direct export to SQL?	Execution time (s)
RegEx + PyMuPDF4LLM	71,43	Yes	41.8
DonutAI/OpenLLaMa-7b	0	No	300+
NotebookLM	38,84	No	67.2

FINAL CONSIDERATIONS AND FUTURE WORK

PyMuPDF4LLM + RegEx:
- High accuracy
- No pre-processing of PDFs needed
- Minimizes non relevant captures

Future work

Cross validate with reference genome

Identify **binding sites**

Score primers based on parameters such as Percentage of Identical Sites and Pairwise Identity to determine **best primer pairs**

Continue **sequence** and **phylogenetic** analysis on the retrieved genomes

Structural analysis of variants will form the foundation for **3D modelling** and **docking** simulations

Acknowledgments:
This research was supported by Portuguese national funds through the Foundation for Science and Technology (FCT) within the scope of UIDB/04423/2020 (CIIMAR) and UIDP/04423/2020 (CIIMAR), 10.54499/LA/P/0008/2020,10.54499/UIDP/50006/2020, UIDB/04050/2020 (UMinho CBMA - https://doi.org/10.54499/UIDB/04050/2020), 10.54499/UIDB/50006/2020 (LAVQ), UIDB/00319/2020 (ALGORITMI/LASI) and UIDB/00127/2020 (IEETA/LASI - doi.org/10.54499/UIDB/00127/2020). JC acknowledges the FCT funding for his research contract established under the transitional rule of Decree Law 57/2016, amended by Law 57/2017. D.P. is funded by national funds through FCT–Fundação para a Ciência e a Tecnologia, I.P., under the Scientific Employment Stimulus—Institutional Call—reference CEECINST/00026/2018.