U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

An automatic algorithm to design primers for identification of highly infectious viruses using inter and intra-specific genome conservation scores

André Rafael Ferreira

Mestrado em Genética Forense

Departamento de Biologia

2021

Orientador

João Carneiro, CIIMAR, Matosinhos, Portugal

Coorientadores

Filipe Pereira, IDENTIFICA Genetic Testing, Maia, Portugal

Luísa Azevedo, Investigadora IPATIMUP/i3S e Professora auxiliar convidada da FCUP

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto,_____/_____/_____

Dissertação de candidatura ao grau de Mestre em Genética Forense, submetida à Faculdade de Ciências da Universidade do Porto.

Dissertation for applying for a Master's degree in Forensic Genetics, submitted to the Faculty of Sciences of the University of Porto.

“Imagination is the real and eternal world of which this vegetable universe is but a faint shadow.”

William Blake

Acknowledgments

I would like to thank my supervisors Dr. João Carneiro, Dra Luísa Azevedo and Dr Filipe Pereira for all the help, support and patience through all this project. Thank you for embracing this adventure and for encouraging me to keep going through all the challenges. To all my friends that always put a smile on my face.

Resumo

Apesar do extraordinário progresso na área das ciências biomédicas, o conhecimento e entendimento sobre a diversidade de zoonoses e patógenos é ainda muito limitado. Enquanto os vírus de DNA tem evoluído por milhões de anos, os vírus de RNA parecem ter uma evolução muito mais recente. Devido à sua polymerase/reverse-transcriptase sujeita a erros (≈ 10-4/site/por ciclo de replicação), vírus de RNA possuem populações muito diversificadas geneticamente. Genomas de RNA são muito mais variáveis em sequência e diferentes variantes genéticas pertencentes ao mesmo vírus normalmente possuem vários polimorfismos.

Neste trabalho pretendemos caracterizar a diversidade genética dos vírus SARS-CoV-2, HIV, Ébola e Influenza, e estabelecer uma base de dados dos melhores oligonucleotídeos para a identificação destes vírus, baseado na tecnologia de PCR. Nós reunimos a informação sobre a sequência de referência e diferentes variantes genéticas para estes vírus a partir de bases de dados online. Posteriormente, realizamos alinhamentos do genoma completo. Desenvolvemos um algoritmo que analisa os alinhamentos obtidos, de forma a reunirmos as zonas mais conservadas de cada vírus. Essas regiões são depois utilizadas para gerar oligonucleotídeos para a identificação dos diferentes vírus, que em termos teóricos serão muito eficientes. O objetivo principal deste trabalho é obter uma base de dados de oligonucleotídeos que possam ser utilizados para ajudar investigadores e trabalhadores da saúde a diagnosticar e identificar rapidamente e com facilidade, estes agentes infeciosos potencialmente perigosos para a população humana.

Palavras-Chave: Vírus RNA; SARS-CoV-2; HIV; Ebola; Influenza; Oligonucleotídeos; Primers; PCR

Abstract

Even with the extraordinary progress in biomedical sciences, the knowledge and understanding of zoonoses and pathogens diversity is still very limited. While DNA viruses have been evolving for millions of years, RNA viruses seem to have a much more recent evolution. Due to their error prone polymerase/reverse-transcriptase (≈ 10-4/site/replication cycle), RNA viruses have the most genetically diversified populations. RNA virus genomes are very variable in sequence and different strains from the same virus usually have many single nucleotide polymorphisms.

In this work we intend to review and characterize the genetic diversity of the SARS-CoV-2, HIV, Ebola and Influenza viruses and establish a database of the best oligonucleotides primers for the identification of the virus based on the PCR technology. We retrieved the information on the complete reference genome and different strains for these viruses from online databases and performed multiple sequence alignments (MAS). We developed an algorithm to analyze these genome alignments, in order to retrieve the most conserved regions for each virus. Those regions were then used to generate the theoretically more efficient oligonucleotide primers for the identification of the different virus. The final objective of this work was to obtain a database of oligonucleotides that can be used to help researchers and health workers diagnose, quickly and with ease, these infectious and potentially dangerous agents for human populations.

Keywords: RNA viruses; SARS-CoV-2; HIV; Ebola; Influenza; Oligonucleotide Primers; Alignments; PCR

Contents

Tables

Figures

List of abbreviations

DNA – Deoxyribonucleic Acid

RNA – Ribonucleic Acid

HGT – Horizontal Gene Transfer

SARS-CoV-2 – Severe Acute Respiratory Syndrome Coronavirus 2

HIV – Human Immunodeficiency virus

RT-PCR – Reverse Transcription Polymerase Chain Reaction

RDTs – Rapid Diagnostic Tests

IATs - Isothermal Amplification Technologies

LAMP - loop-mediated isothermal amplification

dNTPs – Deoxynucleotide triphosphates

nt - Nucleotides

1. Introduction

1.1. Virus

Viruses are microorganisms constituted by very small infectious elements, whose size is measured in millimicrons: mícron – one-millionth of a millimeter. The virion is an entire virus particle consisting of an outer protein shell called capsid and an inner core of nucleic acid, of DNA or RNA, carrying the hereditary information of that organism. Viruses are pervasive companions of cellular life forms, and it seems that every biological entity has its own set of viruses or, at the very least, selfish genetic components that act like viruses. Viruses actively migrate between biomes and are important evolutionary agents due to their ability to act as horizontal gene transfer vehicles (HGT) (Chen et al., 2016).

Viruses are notable for the diversity of their genetic cycles, which contrasts with the consistency of the cellular genetic cycle. Viruses with different genome methods have a wide range of genome sizes and a non-uniform distribution among host species (Koonin et al., 2006).

About 80% of known viruses survive in non-human reservoirs, predominantly farm mammals and poultry, and to a lesser extent, wild animals and arthropods. We have limited information of such zoonosis and the diversity of these viruses in their reservoirs, with data on certain domestic mammals hosting dozens of virus species limited and knowledge of wild animal viruses poor (Parvez & Parveen, 2017).

New outbreaks of infections have resulted in the identification of a varied array of extremely pathogenic viruses, primarily from the Filoviridae, Arenaviridae, Hepeviridae, Coronaviridae, Togaviridae, Bunyaviridae and Paramyxoviridae families. Despite progress in understanding the nature and biology of many harmful viruses, information of emerging viruses is limited. As a result, the ever-present nature of infectious illness emergence and transmission provides an ongoing issue (Parvez & Parveen, 2017).

When healthcare spending in the economically developed world has been constrained, viral outbreaks have had a significant impact on both local and national resources. In poorer regions, where many of these diseases originate, capacity to diagnose and control emerging diseases is severely constrained (Howard & Fletcher, 2012).

1.2. DNA and RNA viruses

The separation between viruses that frequently jump species boundaries and viruses that are carried vertically between species over a long period of evolutionary time is the most simple and basic approach to the evolutionary history of human viruses (Holmes, 2008). In consideration of each virus's mechanism, DNA and RNA viruses tend to mimic this split. DNA viruses are more likely to cause chronic infections in their hosts and to have developed through a long virus-host codivergence process, whereas RNA viruses are more likely to cause just acute infections in their hosts and to have evolved by cross-species transmission. Viruses like HIV and SARS employ this technique to migrate from one host to another (Cleaveland et al., 2001).

Other key factors of viral pathogenicity are the mode of transmission, the rates of evolutionary mutations and virulence. DNA viruses are, more frequently than not, transmitted vertically or sexually, and have lesser or delayed virulence. This is due to the length of time the pathogen must remain in the host in order to assure transmission, requiring more time to evolve. Body fluids, feces, aerosols, and vectors of transmission are all common ways for RNA viruses to spread horizontally. These properties influence virus strength, resulting in high virulence and rapid development. RNA viruses, on the other hand, require huge and well-connected host populations to thrive; any decrease in the number of vulnerable hosts causes the virus to become extinct in the population (Holmes, 2008). The majority of our understanding of RNA virus biodiversity and evolution, as well as their diverse genomic architectures, comes from those that can be cultivated and operate as disease agents in people or economically important animals and plants. Nonetheless, they only account for a small portion of eukaryotic diversity (Shi et al., 2016).

It's plausible to believe that their divergent mutation rates are at the basis of this division. RNA viruses have a high intrinsic mutation rate, estimated to be up to one mutation per genome, each replication, due to replication with either RNA-dependent RNA polymerase (RNA viruses) or reverse transcriptase (retroviruses), neither of which has to ability to proofread or fix errors. (Holmes, 2008; Simon-Loriere & Holmes, 2011). However, while DNA viruses are thought to have evolved and diversified over millions of years, most RNA viruses are thought to have evolved recently and exist as more genetically diversified populations, but only a fraction of the 158 known human RNA virus species have adapted to humans, but roughly 87 percent of the 91 DNA viral species have fully adapted to human hosts. The formation of stable virus lineages in human populations may be aided by viral genetic changes, re-assortment, or virus-host genetic recombination throughout the human adaptation process. As a result, it's feasible that such human-adapted viruses could spread asymptomatically and go unnoticed until their new clinical manifestations are discovered (Parvez & Parveen, 2017).

1.3. SARS-CoV-2 – Severe Acute Respiratory Syndrome

The Coronaviridae family includes viruses with a positive sense, single-strand RNA genome. These viruses have been identified in avian and mammal hosts, including humans. Coronaviruses have genomes from 26.4 kilo base-pairs (kbps) to 31.7 kbps, being the largest among RNA viruses. In December 2019, many cases of pneumonia were successively reported in some hospitals in Wuhan, China. SARS-CoV-2 was later identified as the virus responsible for the acute respiratory infections. So far, the disease has rapidly spread from Wuhan to other areas of China, and it is now reported in 221 countries (Wu et al., 2020).

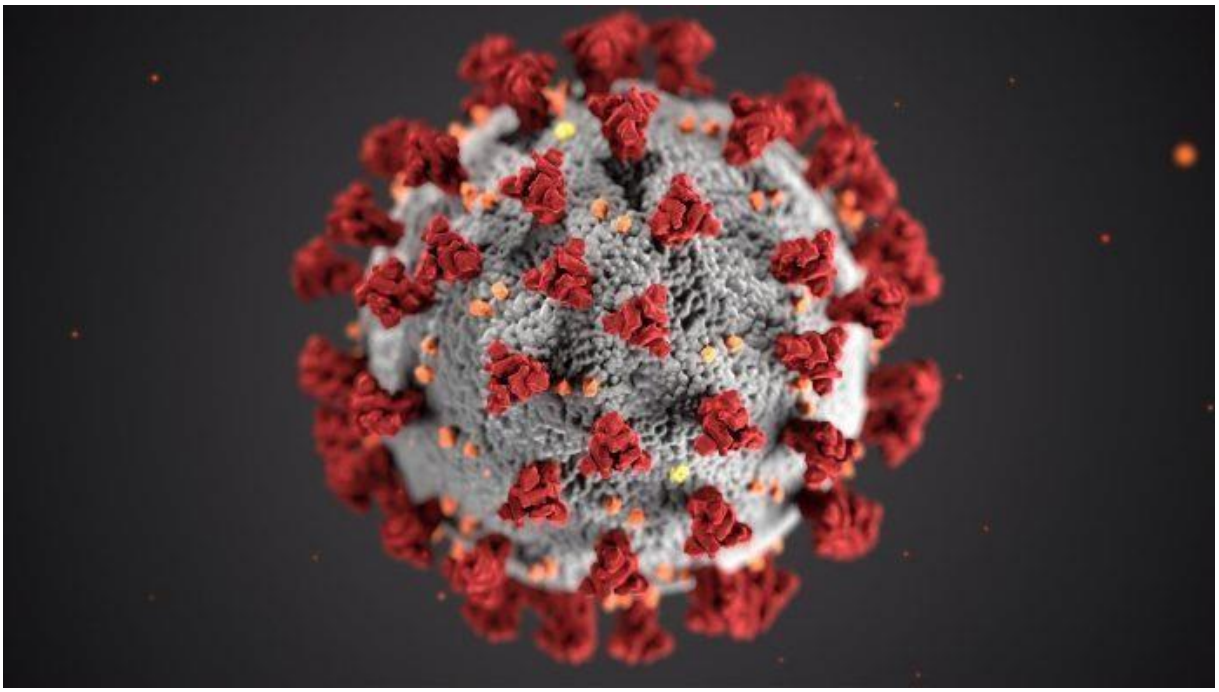


Figure 1 – Computerized reprodution of SARS-CoV-2 (Foto: CDC / Unsplash)

The SARS-CoV-2 outbreak highlights the wild zoonotic reservoir of deadly viruses. Throughout genetic analysis, bats were suggested as native host of SARS-CoV-2, as several evidences showed the homology of the ACE2 receptor and the presence of single intact ORF on gene 8 (Dhama et al., 2020).

Following infection to a human host, this virus is transmitted among the human population by close person-to-person contact, and was also determined that it could be transmitted through air (airborne spread) (Kirtipal et al., 2020).

The rapid detection of SARS-Cov2 remains a crucial part of containment and mitigation strategies. Real-time reverse transcription-PCR (RT-PCR) remains the most common method. However, the demand for increased testing and shortage of RT-PCR reagents, prompt the exploration of alternative testing options, such as rapid diagnostic tests (RDTs) and isothermal amplification technologies (IATs) (Safiabadi Tali et al., 2021).

1.4. HIV - Human Immunodeficiency virus

The Human Immunodeficiency Virus (HIV) epidemic is one of the most devastating in human history. According to the World Health Organization (WHO) since the beginning of the epidemic, 75 million people have been infected with HIV and about 32 million people have died from the infection. HIV is a lentivirus, a smaller group of the Retroviridae family, recognized in the early 90s as the cause of acquired immunodeficiency syndrome (AIDS). Lentivirus group encompasses a wide range of different viruses that infect a diverse group of animal species. HIV genomic size is about 9.8 kb with open reading frames coding for several viral proteins, and the primary transcript is a full-length viral mRNA. There are two types of HIV, HIV-1, and HIV-2, differing genetically by more than 55% of their genomes, with HIV-1 being less widespread (Beloukas et al., 2016).

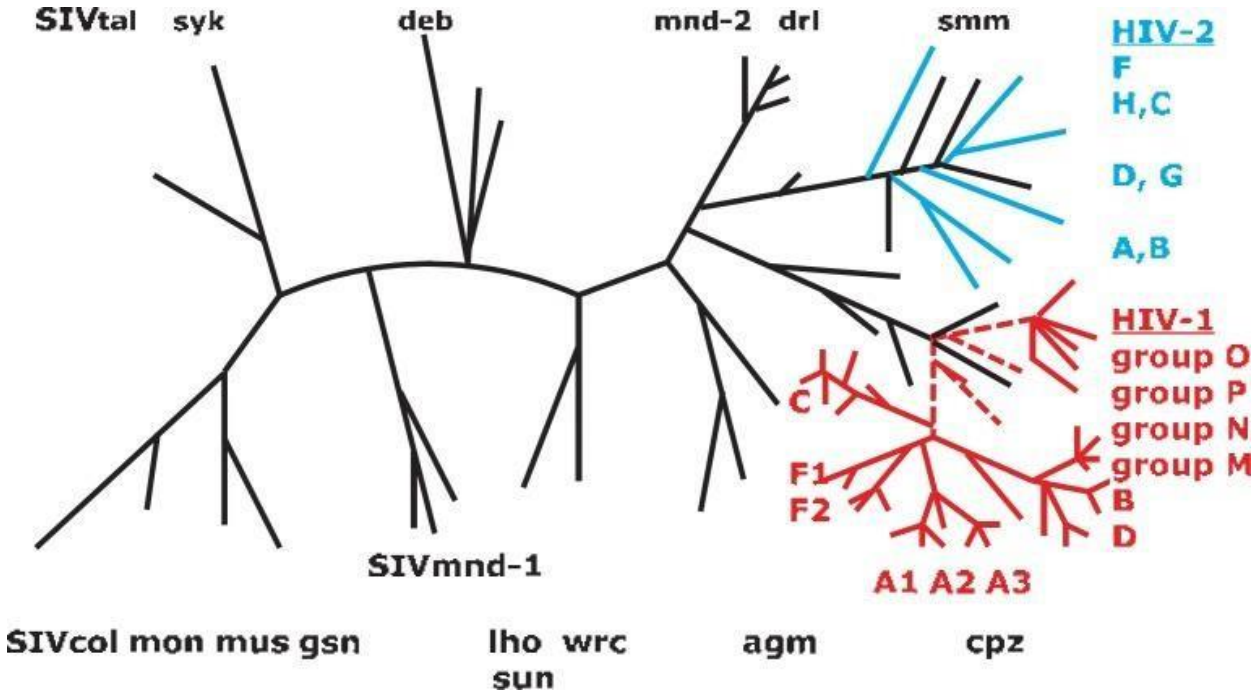


Figure 2 – Scheme of the phylogenetic tree of human and simian Lentiviruses. In red HIV-1 is displayed (Seitz, 2016).

HIV causes a chronic infection that leads to a progressive disease. People with this infection, usually develop AIDS within 10 years of infection, resulting into substantial morbidity and premature death (Prejean et al., 2011).

In spite of the vast array of HIV diagnostic tests, there is still a great demand for novel diagnostic methods that are rapid, cost-effective and capable of detecting recent or acute infections. Nucleic acid-based tests are currently the most used methods. These approaches, however, are limited to laboratory setting, as they are expensive, time-consuming and require laboratory training and infrastructures. The loop-mediated isothermal amplification (LAMP) is a method that have key characteristics ideal for the rapid, cost-effective amplification of nucleic acids, making it an alternative for conventional PCR (Bartholomeusz & Locarnini, 2006).

1.5. Ebola virus disease (EVD)

Ebola virus disease (EVD) is a severe and often lethal disease caused by Ebola virus (EBOV). Outbreaks of this virus consistently start from a single case of zoonotic transmission, followed by person-to-person transmission, via direct contact or contact with infected body fluids. The main symptoms are high fever, gastrointestinal signs and multiple organ dysfunction (Jacob et al., 2020). This virus belongs to the *Filoviridae* family and has a negative single-strand RNA genome. Recently, in 2014, an Ebola outbreak was declared the most widely spread deadly epidemic at the time, showing the world the problem with the outburst spread of deadly viruses and inciting the study on rapid diagnostics, detection and therapeutics for the disease (Kaushik et al., 2020). The World Health Organization (WHO) declared the epidemic a public health emergency of international concern with severe global economic burden (Kaushik et al., 2020). Diagnosis can be achieved through real-time reverse transcription PCR to detect viral RNA or rapid diagnostic tests based on immunoassays to detect EBOV antigens(Jacob et al., 2020).

Figure 3 – Ebola virus (EBOV) (Patel PR, Shah Su, 2021 Jul 21)

1.6. Influenza

The Influenza virus infection is an epidemic disease that affects the world seasonally. Due to that fact, it is of extreme importance for human health, being associated with high morbidity and may lead to serious complications such as viral or bacterial pneumonia, resulting in life-threatening consequences for patients. Older adults, young children, people with chronic diseases, and immunosuppressed people are at higher risk for complications and death from influenza virus infections (Yildirim et al., 2017).

These viruses belong to the family of Orthomyxoviridae and their genome consists of multiple segments of single-stranded negative-sense RNA. Two different strains of this virus are especially important, Influenza A and B since they cause substantial morbidity and mortality in humans. Influenza C viruses can cause sporadic outbreaks, but only mild respiratory disease, and affects mainly children (Blümel et al., 2009). Even though vaccines against influenza A and B are available and produced at large scale, the protection that they offer is limited, due to antigenic variation in the haemagglutinin (HA) and neuraminidase (NA) envelope glycoproteins of these strains. Therefore, emerging antiviral resistance is a continuing challenge (Te Velthuis & Fodor, 2016).

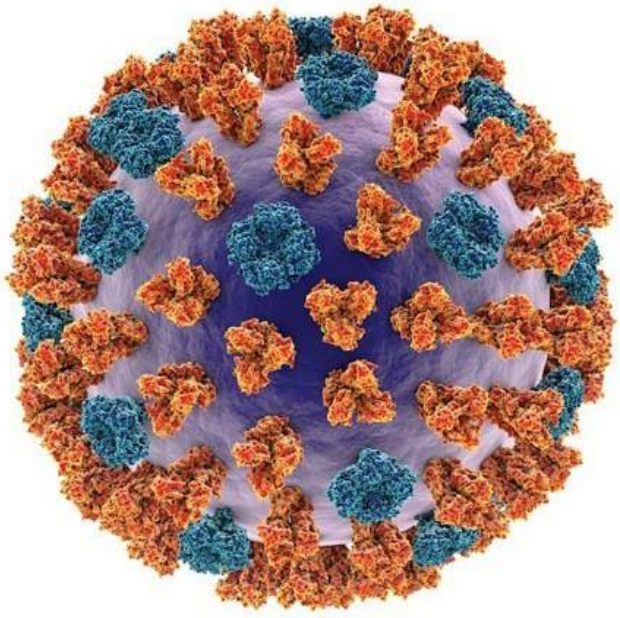


Figure 4 – Influenza virus (Credit: Kateryna Kon/Shutterstock)

The detection methods for Influenza that, for now, dominate the market are RT-PCR and ELISA, but there are many others. We have the classical methods like Rapid

Influenza Diagnostic Tests (RIDTs), immunofluorescence assays, serological assays and nucleic acid-based tests (NATs) or Next-Generation Sequencing. But these methods have some limitations in sensitivity or are of great monetary cost. New approaches have been developed to decrease time of analysis and costs, and to increase sensitivity and efficiency (e.g., microchip approaches and modifications of standard models like LAMP) (Dziabowska et al., 2018).

1.7. Diagnostic and detection methods for viruses

Polymerase Chain Reaction (PCR)

The polymerase chain reaction (PCR) is a powerful tool for the amplification of nucleic acids. The development of PCR-based detection technologies laid the groundwork for the quick and accurate identification of viral nucleic acids in clinical settings (Watzinger et al., 2006). It is a common laboratory technique used to make millions of copies of a genome DNA region. PCR requires a DNA polymerase enzyme that makes new strands of DNA. The DNA polymerase generally used in PCR is called Taq polymerase. Taq polymerase can only make new strands of DNA when a pair of primers is given to the PCR reaction. Primers are short single-stranded artificial DNA (oligomers) about 18 to 25 nucleotides long that anneal to a specific region in the DNA template by complementary base pairing. PCR is used in many different areas of biology and medicine including molecular biology research, medical diagnostics and even in some branches of ecology. Choosing the right primer for PCR is one of the most crucial factors determining the outcome and quality of the PCR (Untergasser et al., 2012)

Aside from specificity check, several other factors are considered for designing a good primer such as the GC content, melting temperature, secondary structures, etc. Temperature cycling is used in PCR to start and stop bursts of enzymes-catalyzed DNA synthesis. There are three stages to each cycle:

- Denaturation of the template DNA using heat (usually >90°C)

- Annealing of two synthetic oligonucleotide primers to the denatured template DNA. These primers, usually 18–25 nucleotides in length, are designed using preexisting knowledge of the DNA sequence of the template. The two primers are complementary to sequences on opposite strands of the target DNA.

- Extension, in which DNA synthesis is initiated at the 3′ ends of the bound primers. Extension of the primers occurs at temperatures between ≈55°C and 70°C in an enzymatic reaction catalyzed by a thermostable DNA polymerase.

This process is done approximately 25–35 times in a thermal cycler, which is a programmed device that controls the time and temperature of each cycle step. The first round of synthesis yields two daughter DNA strands, which serve as templates for the second round of primer-driven DNA synthesis, which generates new DNA chains with a length equal to the number of nucleotides between the binding sites of the two primers 5' ends (Kubista et al., 2006).

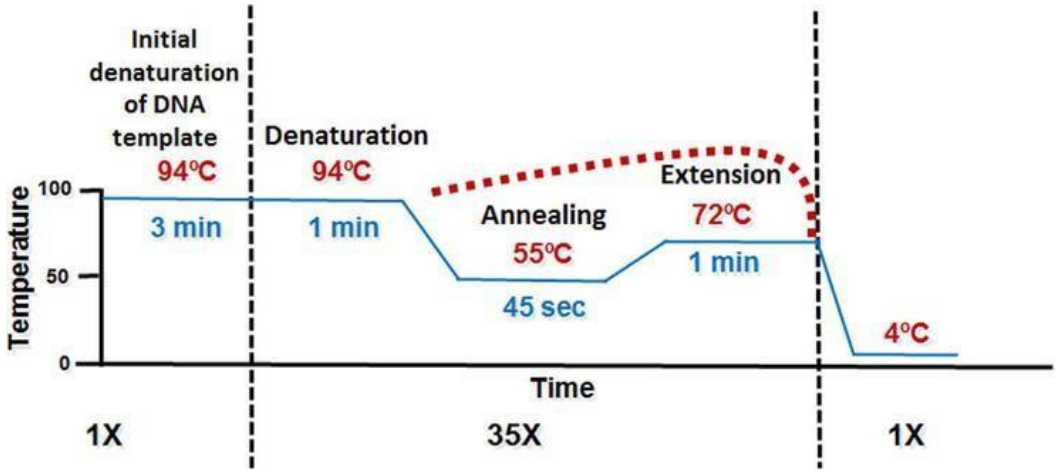


Figure 5 – Polymerase Chain Reaction (PCR) thermocycling protocol (Chang, 1994).

The PCR continues for at least 25 cycles after that, with copies of the target sequence doubling every cycle, until the concentration of primers and/or deoxynucleotide triphosphates (dNTPs) becomes limiting. In actuality, the likelihood of a target molecule being copied in a given cycle is about 1. The presence of inhibitors in the reaction, the features of the thermostable polymerase employed to catalize the PCR, the use of partially degraded template DNA, and mispriming at ectopic places in the template DNA can all cause PCRs to fail to follow perfect kinetics (Green & Sambrook, 2019).

In addition to measuring viral load at a specific time point, quantitative PCR assays can also be used to determine the dynamics of virus growth, monitor treatment response, and distinguish between latent and active infection in viruses that survive in designated cell types. Robust detection and identification are required to fully comprehend the viral illness (Watzinger et al., 2006). RT-PCR is a particular method used for RNA viruses (Zheng et al., 2008). The introduction of RT-PCR allowed for the sensitive detection and analysis of viral activity, allowing for a precise assessment of pathogenetic stages and illness progression in most viral infections (Clementi et al., 1993). This method couples a reverse transcription reaction, based onPCR technology, to generate complementary DNA (cDNA) from mRNA. The template used for reverse transcriptase is an RNA sequence. This technique is advantageous since it requires a relatively small amount of biological sample (Bachman, 2013).

Although RT-PCR is widely used for rapid virus identification due to its low cost and high sensitivity and specificity, very few online databases resources have compiled PCR primers for RNA viruses (Carneiro & Pereira, 2016) (Carneiro et al., 2017) (Carneiro et al., 2020). RNA virus genomes are very variable in sequence, different isolates from the same virus usually have many single nucleotide polymorphisms (SNPs) (Zheng et al., 2008).

The great genetic variety of RNA viruses makes developing efficient nucleic acid-based tests difficult. Variations in the binding sites of RT-PCR primers, small interfering RNAs (siRNAs), and probes might result in false-negative diagnoses or ineffective therapies (Carneiro & Pereira, 2016).

To layout primer pairs with high sensitivity and selectivity, several computational algorithms have been proposed, and most of them take into account signature primers and ordinary short RNA segments in the targeted molecules. With the help of computational methods, we can evaluate primer constraints in its entirety via an expert fitness function. These approaches do stochastic and iterative searches over the full solution, ensuring that attainable primer sets are found in a finite amount of time, which is ideal for high-throughput primer design (Chuang, 2013).

1.8. Design of oligonucleotide primers for basic PCR

Primers are designed for specificity, which is attained when a primer pair anneals to its target sequence in the template DNA in a stable manner. The better the specificity of an oligonucleotide for a given target, the longer it is. The likelihood that a sequence perfectly complementary to a string of nucleotides will occur by chance within a DNA sequence space containing a random sequence of nucelotides can be calculated using the following equations. G,C,A and T are the number of particular nucleotides on the oligonucleotide, and K is the estimated frequency of occurrence within the sequence space, g is the sequence's relative G+C content.

$$K = [g/2]^{G+C} \times [(1-g)/2]^{A+T},$$

Several equations are available to determine the melting temperature of any given oligonucleotide primer and its complementary target sequence. One of the most used equations, known as "The Wallace Rule" takes in account the sum of A and T residues (A+C), and the sum of G and C residues (G+C) in the oligonucleotide. The result is T_m, the melting temperature, expressed in ºC, for that oligonucleotide (Green & Sambrook, 2019).

$$T_m = 2(A+T) + 4(G+C),$$

To be specific and efficient, GC content and melting temperature should be properly matched within a primer pair. A lower melting temperature can result in a loss of specificity, whereas a high melting temperature can result in mispriming (Dieffenbach et al., 1993).

1.8.1. Types of PCR primers

There are several types of primers, therefore choosing the right primer is of utmost importance for the quality of the PCR. The more information we know about the DNA sequence to be analyzed, the easier it will be to choose the right primer pair (Green & Sambrook, 2019). These sequences can be:

Universal primers - are primers that anneal with many different types of DNA templates. Universal primers are commonly found in cloning vectors.

Specific primers - are used in places where the genetic sequence is already known, or a specific gene is a target. Specific primers can be purposefully designed to amplify a particular gene, which increases the success rate of PCR.

Special primers - degenerate primers that have specific uses, such as the amplification of the same gene from two different organisms in which the DNA sequence complementary to the primer sequence is not invariant between the two species under analyses.

1.9. PCR-field of applications

PCR has impacted many diverse fields including medicine, molecular paleontology and forensics. Forensics is one of the most well-known applications of PCR. Before the advent of PCR, a small sample of DNA may not have been sufficient to perform testing and draw an accurate conclusion. However, by using PCR scientists and researchers can now generate enough DNA to identify remains, for example at the scene of accidents and natural (or man-made) disasters such as earthquakes or fires. Similarly, samples of DNA recovered from crime scenes can be replicated using PCR to identify criminals (Watzinger et al., 2006).

Modern medicine has several applications of PCR, for example, PCR has become an important tool in the field of cancer research. It can be used to detect some virally-induced cancers such as cervical cancer caused by the HPV (Human Papilloma Virus). PCR is also used in the field of medicine to diagnose diseases with a genetic component. For instance, if a cancer patient's DNA shows a specific genetic mutation that could indicate whether the individual will respond better to one treatment over another. Another important medical application of PCR is in the procedure of tissue typing for patients who need to undergo an organ transplant (Valones et al., 2009).

Infectious diseases in birds, animals or humans can also be detected and monitored using PCR. PCR is faster than other methods to identify infectious diseases, which allow starting treatment at an early stage. PCR can identify both fungal and parasitic infection and can, for example, detect HIV infection even before any antibodies have been produced (Watzinger et al., 2006).

1.10. Biological Computational Approaches

The importance of algorithms in Molecular Biology Research

With the appearance of automated computing devices, an algorithm has become equivalent with a description that can be turned into a computer program that instructs a computer how to solve the problem addressed by the algorithm. The amazing capacity of computers to carry out billions of single calculations per second, and to store bits of information, makes it possible to address a wide range of problems that would otherwise remain out of reach. One field of application for these algorithms is an interdisciplinary field called Computational Biology, which is focused with making use of the capacities of computers to face problems of biological interest. The activities range from algorithmic theory focusing on problems with biological relevance, via construction of computational tools for specific biological problems, to experimental work, making use of computational tools written to analyze huge amounts of biological data. The quality of an algorithm is a combination of its running time and space assumption, plus the biological relevance of the answers produced. These parameters both depend on the modeling of the biological reality that led to the formulation of computational problem that is addressed. The set-up of a good algorithm that addresses a specific problem is therefore an interdisciplinary work that involves interchanging between modeling the biological context and constructing the algorithm, until a sensible balance between the running time, space assumption and biological relevance of the answers produced is accomplished (Jiang & Feng, 2013).

For projects involving PCR amplification and/or DNA sequencing, proper primer design is critical, and a large number of algorithms and design guidelines have been proposed. Prior to beginning a project, in silico design may help to avoid difficulties in trials and analysis. Design is often done in separate, non-communicating processes, in part because it is complicated, even for well-defined objectives with little or no genetic diversity. Because limitations in one phase can affect another, optimizing the overall design when working in a sequence of independent steps can be difficult or impossible. As a result, a strategy that can optimize primers and probes while simultaneously addressing all design requirements is required (Brodin et al., 2013).

2. Methods

The main goal of this study was to develop an algorithm that would incorporate the automatic retrieving of new genomic sequences from different sources for the design of accurate primer sequences of distinct viral organisms. For this we created an algorithm to read the genomic sequences.

For the design of the PCR or RT-PCR primer set, a reference sequence for the viral genome was used which comprises the targeted regions for amplification. This is a key factor to guarantee that the primer pair does not amplify non-specific products.

The sequences, including strains and the reference sequence, were retrieved from online databases. The sequences were submitted to an alignment process to analyze the variation at different genomic positions so that the most conserved regions were well established. The genomic conservation was calculated using percentage of identical sites (PIS) and percentage of pairwise identity (PPI). The top conserved regions were used to design the most accurate primers considering primer length (optimal length between 18-25 nucleotides (nt)), primer melting temperature (50-80ºC), GC content (40-60% the percentage of total nt).

2.1. Retrieving of SARS-CoV-2, HIV, Ebola and Influenza sequences

We started by retrieving the nucleotide sequences of the reference genomes and strains for SARS-CoV-2, HIV, Ebola and Influenza viruses:

a) SARS-CoV-2 virus – For this virus we used the GISAID Initiative (https://www.gisaid.org/) that promotes the rapid sharing of data from all coronavirus causing COVID-19. This includes genetic sequence and related clinical and epidemiological data associated with human viruses, and geographical as well as species-specific data associated with avian and other animal viruses, to help researchers understand how viruses evolve and spread during epidemics and pandemics.
COVID CG (https://covidcg.org/) is an open resource for tracking SARS-CoV-2 single nucleotide variations (SNVs), lineages and clades using the virus genomes on the GISAID database while filtering by location, date, gene, and mutation of interest. We used the COVID CG to confirm all the variants detected in SARS-CoV-2 genome.

b) HIV virus - For HIV, we avail oneself of the HIV database (https://www.hiv.lanl.gov/), that contains comprehensive data on HIV genetic sequences. It also provides access to a set of tools that can be used to visualize these data. We used pre-build alignments from http://portugene.com/HIVoligoDB/cgi-bin/HIVoligoDB (Carneiro et al., 2017), to compare with the previously obtained sequences. A curated final alignment was generated for HIV-1 and HIV-2.

c) Ebola virus - We retrieved the nucleotide sequences of Ebola virus disease from EbolaID Database Project (https://ebolaid.portugene.com/). EbolaID is a free database designed to clear the way for the design of accurate molecular methods for detection and identification of the Ebola virus. It provides an interface for searching, filtering and downloading data from published oligonucleotide sequences annotated according to a reference genome (Carneiro & Pereira, 2016).

d) Influenza virus - Lastly, the sequences for Influenza virus were retrieved from Influenza Research Database (https://fludb.org/) considering the virus type. The main goal of the Influenza Research Database (IRD) is to provide a resource for the influenza virus research community that will facilitate an understanding of the influenza virus and how it interacts with the host organism, leading to new treatments and preventive actions. Due to the enormous amount of data on Influenza and the limited computational capacity for the analysis, we only analyzed the PB2 Segment.

2.2. Filtering and Data treatment

SARS-CoV-2, HIV, Ebola and Influenza Alignments

We used the complete or partial genomic sequences retrieved from the online platforms to perform the alignment as follows:

a) For SARS-CoV-2 we used the MAFFT software, from CIPRES platform, to perform alignments.

b) For HIV, we used a pre-made alignment, retrieved from (https://portugene.com/HIVoligoDB/cgi-bin/HIVoligoDB/). The alignment was then refined with MUSCLE (Edgar, 2004) with Geneious Prime 2021.1.1 (https://www.geneious.com/).

c) For Ebola, the pre-made alignments were obtained from (https://ebolaid.portugene.com/cgi-bin/EbolaID/). The alignment was then refined with MUSCLE (Edgar, 2004) with Geneious Prime 2021.1.1 (https://www.geneious.com/).

d) For Influenza the pre-made alignments were obtained from https://fludb.org/, considering the sequence for the largest genomic segment PB2. We filtered the sequences for human host and excluded the laboratory strains. The alignment was then refined with MUSCLE (Edgar, 2004) with Geneious Prime 2021.1.1 (https://www.geneious.com/).

Alignment's curation

We applied the following filters to the alignment sequences using Biopython (e.g., BioSeqIO module that provides a uniform interface to input and output sequence file formats, including multiple sequence alignments):

a) Alignments with sequences that shared exactly the same nucleotides were eliminated, and only one of them was considered to the analysis.
b) The sequences that have a length below 98% of the reference genomic sequence (seq_len>(0.98*length)) were eliminated as well. The others were saved in a filtered file with the remaining sequences. This was achieved by programming different functions (e.g., the function in Python represented in figure 6 goes through the genome sequence and calculates the genome length of each sequence.
c) After that, the algorithm reads the filtered alignment and eliminate sequences with 1 or more ambiguous base pairs (figure 7). The purpose is to only use sequences with complete genomes (SARS-CoV-2, HIV, Ebola) or partial genome full segments of the genomes (Influenza).

```
def filter_seqs_by_length(self,file,length):
    seqs_to_save=[]
    file_to_save=open("InfluenzaFil.fasta", "w+")
    for sequence in SeqIO.parse(file, "fasta"):
        seq_name=sequence.id
        seq=sequence.seq
        seq_len=len(seq)
        if seq_len >(0.98*length):
            seqs_to_save.append([str(seq_name),str(seq)])
    for sequences in seqs_to_save:
        seq_name=sequences[0]
        seq=sequences[1]
        print (seq_name+"\n")
        file_to_save.write(">"+seq_name+"\n")
        file_to_save.write(seq+"\n")
```

Figure 6 – Function for filtering the nucleotide sequences by length

```
def eliminate_seqs_with_Ns_CoV2ID_alignments(self, file):
    seqs_to_save=[]
    file_to_save=open("InfluenzaFinal.fasta", "w+")
    for sequence in SeqIO.parse(file, "fasta"):
        seq_name=sequence.id
        seq=sequence.seq
        number_N=seq.count("N")
        if number_N==0:
            seqs_to_save.append([str(seq_name),str(seq)])
    for sequences in seqs_to_save:
        seq_name=sequences[0]
        seq=sequences[1]
        print(seq_name+"\n")
        file_to_save.write(">"+seq_name+"\n")
        file_to_save.write(seq+"\n")
```

Figure 7 – Function for filtering the nucleotide sequences for unidentified nucleotides

2.3. Sliding Window Analysis

Sliding window method performs an analysis of the genomic sequence by segments. The results of a sliding window analysis can be resumed using a plot where the sliding window position is represented in the X axis, and the observed calculated measure of interest is represented in the Y axis. This type of analysis is employed to study the properties of different types of genomic sequences (e.g., RNA viruses, human genes, mitochondrial DNA), and different measures (e.g., conservation scores). A sliding window analysis was made using the final SARS-CoV-2, HIV and Ebola viruses complete genome alignments and the largest genome segment (PB2) of the Influenza virus. We analyzed the genomic data using windows with lengths of 50, 100, 200 and 300 nucleotides (nt). These different lengths were used to detect small region-specific variations of the conservation measures (windows lengths <= 100), and global region-specific variation of the conservation measures (windows lengths >=200). Afterwards we processed the conservation scores calculations along the alignment of complete genomes sequences using the python algorithm developed. The conservation measures used were the percentage of identical sites (PIS) and the percentage of pairwise identity (PPI).

In this function (Figure 8 and 9) the program reads the input files, considering the first position as a start point, and then create the window, assuming the end point is the start point plus (+) the length of the window. We added a step variable that was by default of one (1), so the program can calculate the window using jumping steps of one nucleotide (e.g., window ranging from 0-100,1-101,2-102, etc..). The results from these calculations are saved as a CSV file, with the window positions and the PIS and PPI values.

```
def calculate_sliding_windows_pis(self,alignment_file,window_len,step):
    alignment_data = AlignIO.read(open(alignment_file), "fasta")
    alignment_len=alignment_data.get_alignment_length()
    general_data = np.empty([1,3])
    columns = ['StartPosition', 'EndPosition', 'PISValue']
    for window_position in range(1,alignment_len-window_len,step):
        start_pos=window_position
        end_pos=start_pos+window_len
        PIS_value=masterAF_genome_analysis.calculate_PIS_window(alignment_file,start_pos,end_pos)
        print ("The percentage of identical sites between position "+str(start_pos)+"-"+str(end_pos)+" is "+str(PIS_value))
        stack_data = np.stack((start_pos,end_pos,PIS_value), axis=0)
        general_data=np.vstack((general_data,stack_data))
    df=pd.DataFrame(data=general_data,columns=columns,index=None)
    df=df.drop([0])
    df=df.astype({'StartPosition':int})
    df=df.astype({'EndPosition':int})
    df.to_csv('PISvalues.csv', sep=';', index=False)
```

Figure 8 – Function for Sliding Windows for percentage of identical sites

```
def calculate_sliding_windows_ppi(self,alignment_file,window_len,step):
    alignment_data = AlignIO.read(open(alignment_file), "fasta")
    alignment_len=alignment_data.get_alignment_length()
    general_data = np.empty([1,3])
    columns = ['StartPosition', 'EndPosition', 'PPIValue']
    for window_position in range(1,alignment_len-window_len,step):
        start_pos=window_position
        end_pos=start_pos+window_len
        PPI_value=masterAF_genome_analysis.calculate_PPI_window(alignment_file,start_pos,end_pos)
        print ("The percentage of pairwise identity between "+str(start_pos)+"-"+str(end_pos)+" is "+str(PPI_value))
        stack_data = np.stack((start_pos,end_pos,PPI_value), axis=0)
        general_data=np.vstack((general_data,stack_data))
    df=pd.DataFrame(data=general_data,columns=columns,index=None)
    df=df.drop([0])
    df=df.astype({'StartPosition':int})
    df=df.astype({'EndPosition':int})
    df.to_csv('PPI_EBOLA_Window.csv', sep=';', index=False)
```

Fig.9 – Function for Sliding Windows for percentage of pairwise identity

We also used pandas, a software library for data manipulation and analysis. It has tools for loading data into in-memory data objects from different file formats and data alignments and integrated handling of missing data.

2.4. Percentage of Identical Sites (PIS) and the Percentage of Pairwise Difference (PPI).

The most conserved regions in the viruses' genomic sequences were ranked considering two main measures of sequence conservation:

- Percentage of Identical Sites (PIS)

The PIS is calculated by dividing the number of equal positions in the alignment for an oligonucleotide by its length.

- Percentage of Pairwise Identity (PPI)

The PPI is calculated by counting the average number of pairwise matches across the positions of the alignment where the oligonucleotide is located. We then divide this value by the total number of pairwise comparisons.

The algorithm creates the sliding window, the user can choose the length, and then will calculate the frequency of each nucleotide per column through all the sequences of the input. After that, is computed the calculation of the percentual value of identical sites (Figure 10) and pairwise identity (Figure 11) for that window.

```
def calculate_PIS_window(self,file,start,end):
    alignment_data = AlignIO.read(open(file), "fasta")
    alignment=alignment_data[:,start:end]
    #print (alignment)
    summary_align = AlignInfo.SummaryInfo(alignment)#Código novoº
    alignment_len=alignment.get_alignment_length()
    alignment_number_seqs=len(alignment)
    dict_column_freqs={}
    Number_identical_sites=0
    #Calculates the freqs per column
    for col in range(alignment_len):
        alignment_columns=summary_align.get_column(col)
        countA=alignment_columns.count("A")/alignment_number_seqs
        countT=alignment_columns.count("T")/alignment_number_seqs
        countG=alignment_columns.count("G")/alignment_number_seqs
        countC=alignment_columns.count("C")/alignment_number_seqs
        countN=alignment_columns.count("N")/alignment_number_seqs
        countDEL=alignment_columns.count("-")/alignment_number_seqs
        dict_column_freqs[col]=(countA,countT,countG,countC,countN,countDEL)
        #print (alignment_columns[0])
        #print (alignment_columns[1])
    #Calculates percentage of identical sites
    for col in dict_column_freqs:
        data_col=dict_column_freqs[col]
        #print (data_col)
        for value in data_col:
            if value==1:
                Number_identical_sites+=1
            else:
                pass
    #print (Number_identical_sites)
    Percentage_identical_sites=(Number_identical_sites/alignment_len)*100
    #print (dict_column_freqs[0])
    print (Percentage_identical_sites)
    print("Alignment length %i" % alignment.get_alignment_length())
    return Percentage_identical_sites
```

Figure 10 – Function for calculation of the percental value for identical sites

```
def calculate_PPI_window(self,file,start,end):
    alignment_data = AlignIO.read(open(file), "fasta")
    alignment=alignment_data[:,start:end]
    #print (alignment)
    summary_align = AlignInfo.SummaryInfo(alignment)
    alignment_len=alignment.get_alignment_length()
    alignment_number_seqs=len(alignment)
    #Calculates percentage of pairwise identity
    for col in range(alignment_len):
        data_col=summary_align.get_column(col)
        #print (data_col)
        n_comparacoes=0
        n_equal=0
        for i in range(len(data_col)):
            value=data_col
            value1=value[i]         #Nucleótido na posição i
            for j in range(len(data_col)):
                value2=value[j]     #Nucleótido na posição j
                if j<i:             #Só comparar nucleótidos em que a posição j seja sempre superior a i
                    if value1==value2:  #Nucleótidos iguais
                        n_comparacoes+=1
                        n_equal+=1
                    if value1!=value2:      #Nucleótidos diferentes
                        n_comparacoes+=1
                        n_equal+=0
    #print (n_comparacoes)
    Percentage_pairwise_identity=(n_equal/n_comparacoes)*100
    #print(Percentage_pairwise_identity)
    return Percentage_pairwise_identity
```

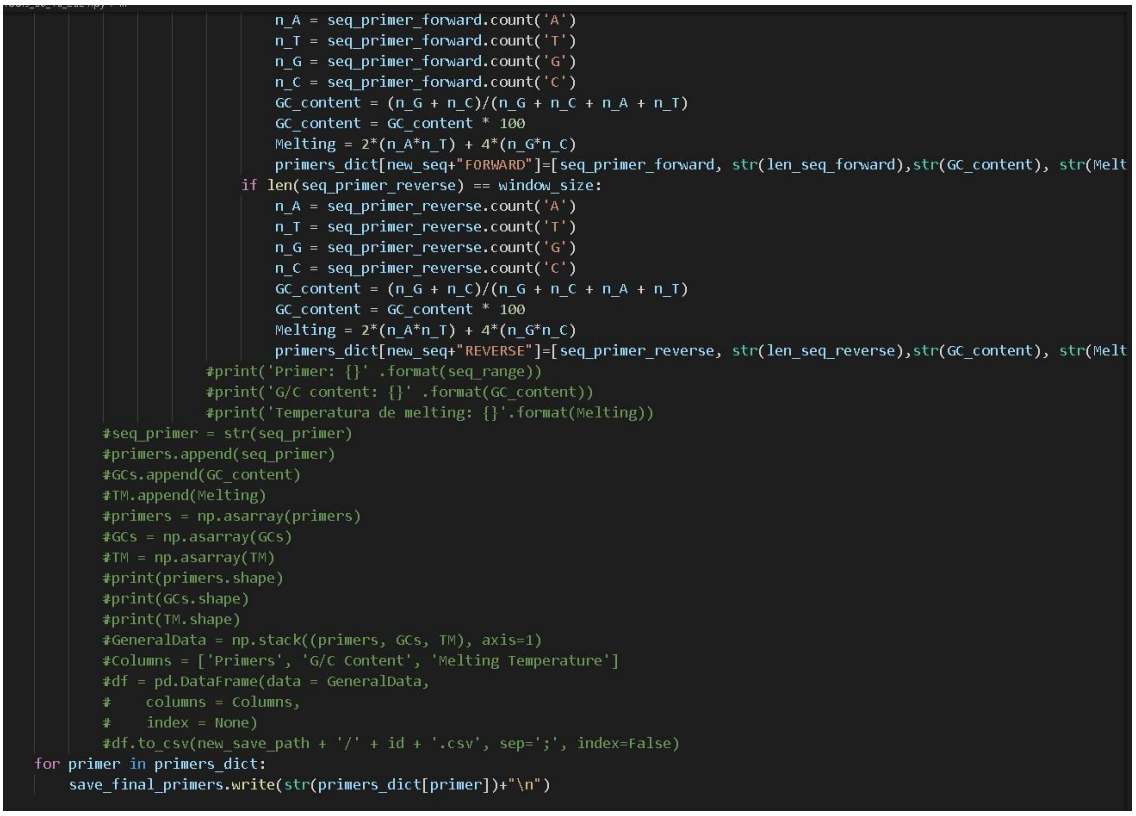
Figure 11 – Function for calculation of the percental value for pairwise identity

2.5. Primer Design

For the primer design, we created an algorithm to go through the nucleotide sequence alignments (Figure 12 and 13), limiting the search on the most conserved regions previously calculated, reading the sequence and writing a list for the possible number of suitable primers. The program reads the RNA sequence for the given input, and writes the complement and the reverse complement DNA sequence for that target. Then it runs calculations for GC content and melting temperature. Finally, it creates a file, with the list for possible forward and reverse primers.

```
def generate_primers(self,file,primer_len,virus):
    alignment_data = AlignIO.read(open(file), "fasta")
    save_final_primers=open("PRIMERS/Final_Primers-"+str(primer_len)+"-"+virus+".csv","w+")
    save_final_primers.write("Primer sequence, Lenght, G/C Content, Melting Temperature, Type")
    primers_dict={}
    #print (alignment_data)
    records = list(SeqIO.parse(open(file),"fasta"))
    df = pd.read_csv("Data_Input/HIV/Conservadas_HIV2_.csv", sep=';', index_col=None)
    StartPoint = df['StartPoint']
    EndPoint = df['EndPoint']
    window_size = primer_len
    for i in range(len(records)):
        seq = records[i].seq
        id_seq = records[i].id
        print(str(id_seq)+"\n")
        primers, GCs, TM = [], [], []
        for j in range(len(StartPoint)):
            new_seq = seq[StartPoint[j]:EndPoint[j]]
            new_seq_str=str(new_seq).replace("-","")
            print (new_seq)
            if len(new_seq_str)==50:
                if new_seq not in primers_dict:
                    complement_seq=new_seq
                    complement_seq=complement_seq.back_transcribe()
                    complement_seq=complement_seq.complement()
                    reverse_seq = new_seq
                    reverse_seq = reverse_seq.back_transcribe()
                    reverse_seq = reverse_seq.reverse_complement()
                    len_seq = len(reverse_seq)
                    for n in range(0,len_seq,1):
                        if n+window_size<=50:
                            seq_primer_forward=complement_seq[n:n+window_size]
                            len_seq_forward=len(seq_primer_forward)
                            seq_primer_forward=seq_primer_forward.str()
                            seq_primer_reverse = reverse_seq[n:n+window_size]
                            len_seq_reverse=len(seq_primer_reverse)
                            seq_primer_reverse=seq_primer_reverse.str()
                            if len(seq_primer_forward) == window_size:
```

Figure 12 – Function for generating possible primers for the most conserved regions



```
                        n_A = seq_primer_forward.count('A')
                        n_T = seq_primer_forward.count('T')
                        n_G = seq_primer_forward.count('G')
                        n_C = seq_primer_forward.count('C')
                        GC_content = (n_G + n_C)/(n_G + n_C + n_A + n_T)
                        GC_content = GC_content * 100
                        Melting = 2*(n_A*n_T) + 4*(n_G*n_C)
                        primers_dict[new_seq+"FORWARD"]=[seq_primer_forward, str(len_seq_forward),str(GC_content), str(Melt
                    if len(seq_primer_reverse) == window_size:
                        n_A = seq_primer_reverse.count('A')
                        n_T = seq_primer_reverse.count('T')
                        n_G = seq_primer_reverse.count('G')
                        n_C = seq_primer_reverse.count('C')
                        GC_content = (n_G + n_C)/(n_G + n_C + n_A + n_T)
                        GC_content = GC_content * 100
                        Melting = 2*(n_A*n_T) + 4*(n_G*n_C)
                        primers_dict[new_seq+"REVERSE"]=[seq_primer_reverse, str(len_seq_reverse),str(GC_content), str(Melt
                    #print('Primer: {}' .format(seq_range))
                    #print('G/C content: {}' .format(GC_content))
                    #print('Temperatura de melting: {}'.format(Melting))
        #seq_primer = str(seq_primer)
        #primers.append(seq_primer)
        #GCs.append(GC_content)
        #TM.append(Melting)
        #primers = np.asarray(primers)
        #GCs = np.asarray(GCs)
        #TM = np.asarray(TM)
        #print(primers.shape)
        #print(GCs.shape)
        #print(TM.shape)
        #GeneralData = np.stack((primers, GCs, TM), axis=1)
        #Columns = ['Primers', 'G/C Content', 'Melting Temperature']
        #df = pd.DataFrame(data = GeneralData,
        #    columns = Columns,
        #    index = None)
        #df.to_csv(new_save_path + '/' + id + '.csv', sep=';', index=False)
for primer in primers_dict:
    save_final_primers.write(str(primers_dict[primer])+"\n")
```

Fig.13 – Function for generating possible primers for the most conserved regions

We then proceeded with selecting the primers that suited best on specific values for GC content and melting temperature. We determined the range for GC content between 38-60%, and the melting temperature between 50-80ºC as established by (Green & Sambrook, 2019).

2.6. Data Input Parameters for Primers Calculations

To calculate the best primers, we used the data as follows:

- Start Point: the first position of the established conserved regions, based on the sliding windows results.
- End Point: the last first position of the established conserved regions, based on the sliding windows results
- PIS value: the percentage of identical sites for that specific region
- PPI value: the percentage of pairwise identity for that specific region
- Weighted average between PIS and PPI: a calculation that takes into account the varying degrees of importance of the numbers in a data set. The calculation established for this value was W=(PPIx5 + PIS)/n, where ‘n’ is the total number of regions on the input.

3. Results and Discussion

3.1. PIS and PPI

3.1.1. SARS-CoV-2

The data for this analysis contained 1,403 sequences for SARS-CoV-2, including the reference sequence. The minimal sequence length was 29,521 and the max sequence length was 29,945. The overall percentage of identical sites of the alignment was 93.6%, and the overall percentage of pairwise identity was 99,95%.

Our results clearly demonstrate the parts of the genome that are most conserved. For the SARS-CoV-2, the alignment showed a high level of conservation, with small variations between the different sequences. We detected the highest values of percentage of identical sites (PIS) (Figure 14) between the 28,000 and 29,903 nucleotide positions (≈90%) in the alignment. These regions contain the nucleocapsid phosphoprotein, N protein, an RNA-binding protein critical for viral genome packaging and viral assembly, and the Open reading frame 10 (ORF10), a unique SARS-CoV-2 accessory protein, which contains eleven cytotoxic T lymphocyte (CTL). The highest values of PIS, including the respective window length were annotated (Table 1).

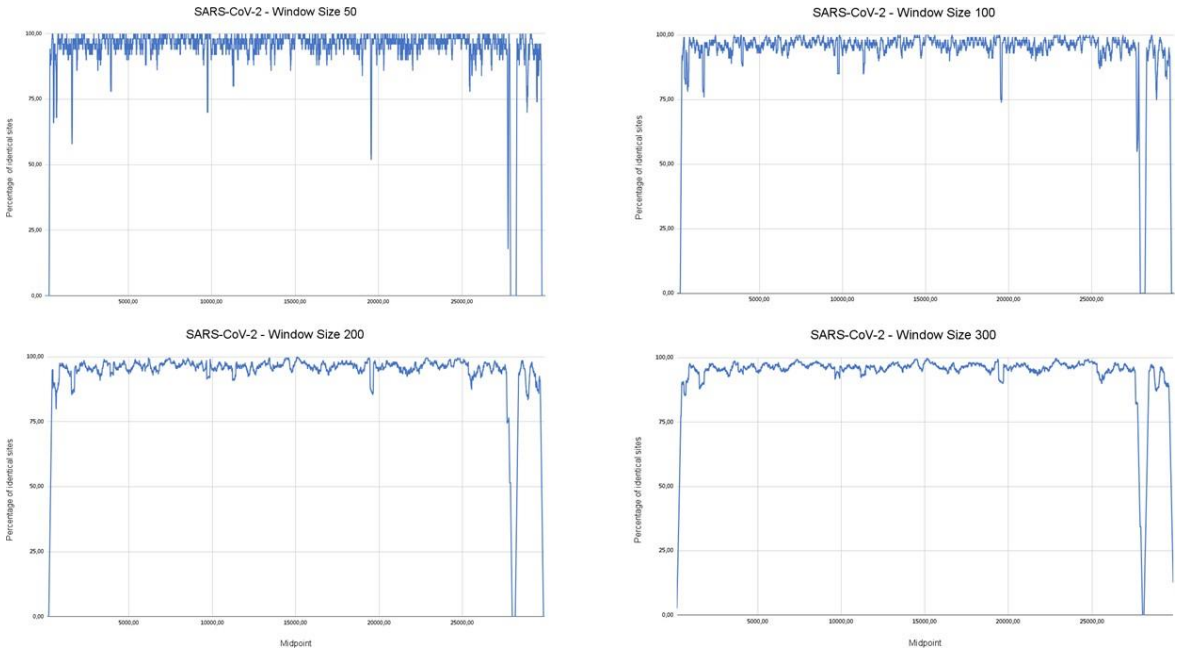


Figure 14 – Percentage of identical sites (PIS) for SARS-CoV-2 genome.

Window length											
50			100			200			300		
Start	End	PIS	Start	End	PIS	Start	End	PIS	Start	End	PIS
28354	28404	92	29345	29445	89	28984	29184	83.5	28987	29287	81
28762	28812	90	29351	29451	88	28977	29177	82.5	28978	29278	80.67
29366	29416	90	29353	29453	87	28991	29191	82.5	28978	29278	80.67
28355	28405	90	29354	29454	86	28964	29164	82	28980	29280	80.67
28762	28812	90	29356	29456	85	28621	28821	81.5	28987	29287	80.67
28363	28413	88	29369	29469	84	28966	29166	81.5	28281	28581	80.33
28769	28819	88	29083	29183	84	28969	29169	81.5	28963	29263	80.33

Table 1 – Established conserved regions for SARS-CoV-2.

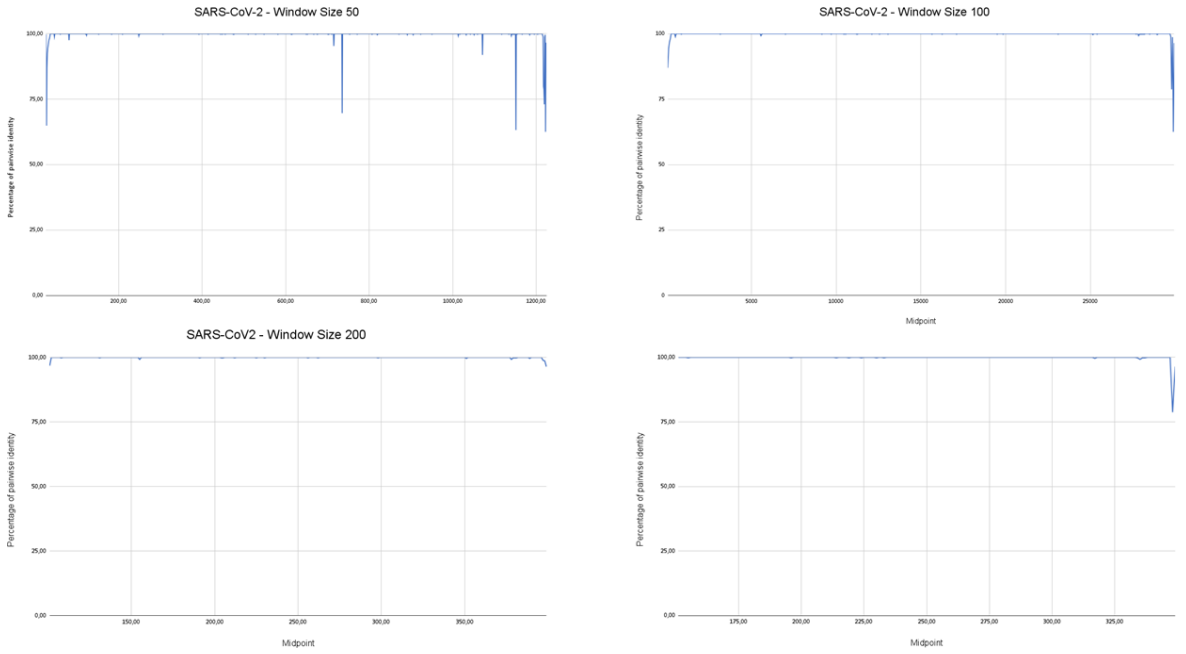


Figure 15 – Percentage of pairwise identity (PPI) for SARS-CoV-2 genome.

3.1.2. HIV-1

For the human immunodeficiency virus, we did a separated analysis for HIV-1 and HIV-2. The data for this analysis contained 4,052 sequences for HIV-1, including the reference sequence. The minimal sequence length was 7,073 and the max sequence length was 9,913. The overall percentage of identical sites of the alignment was 0,6%, and the overall percentage of pairwise identity was 87,2%.

We detected the highest values of percentage of identical sites (PIS) (Figure 16) for HIV-1 between the 3500 and 4100 nucleotide positions (≈ 39%), and 9000 and 9300 nucleotide positions (≈ 98%). The analysis of the genome presented high values of percentage of pairwise identity (Figure 17), with many regions with a value of 100%. The first region contains the *pol* gene, a gene critical for the synthesis and integration of viral DNA into the host genome and the production of capsid proteins. The products of *pol* gene include the HIV-1 reverse transcriptase, an integrase, and the late-phase protease.

These proteins complex with RNase and *vpr* in the cell membrane to form the viral reverse transcription complex. The highest values of PIS and the respective window length were annotated. The second region contains a 3' Long Terminal Repeat (3'-LTR), situated at the end of the HIV genome and is involved in the regulation of viral expression and the initiation and termination of viral RNA transcription. The highest values of PIS and PPI, including the respective window length were annotated (Table 2).

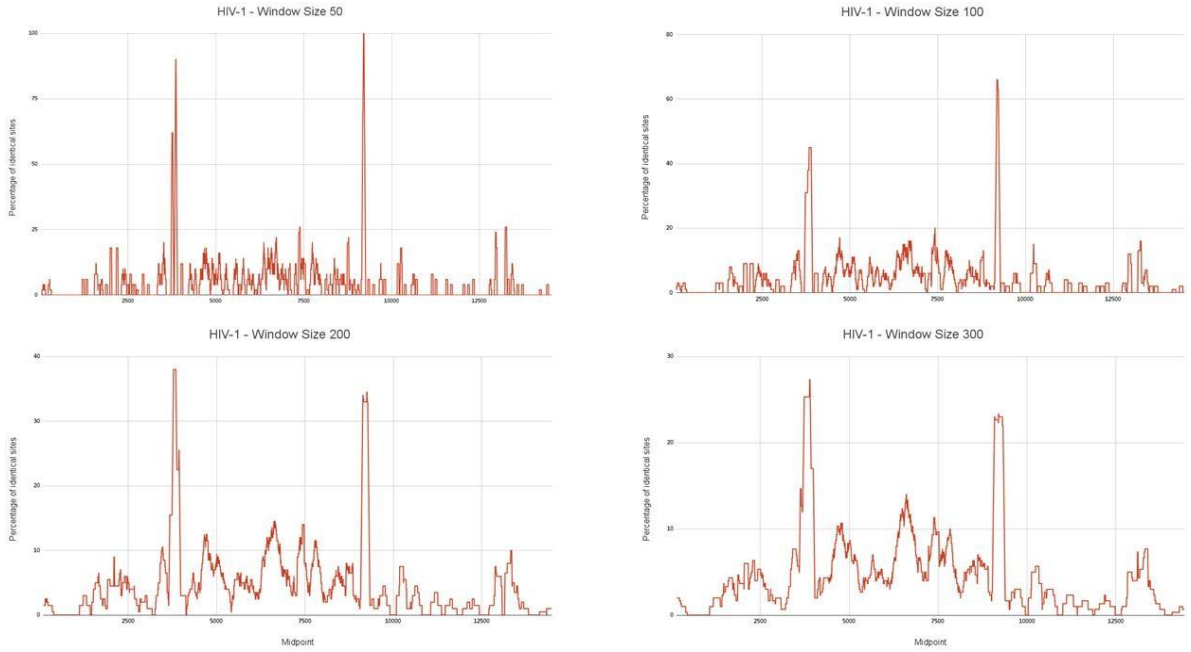


Figure 16 – Percentage of identical sites (PIS) for HIV-1 genome.

Window length											
50			100			200			300		
Start	End	PIS	Start	End	PIS	Start	End	PIS	Start	End	PIS
9174*	9237*	100	9137*	9256*	66	3685*	3947*	38	3738*	4047*	27,33
9173	9223	98	9136	9236	65	3684	3884	37,5	3737	4037	27
9188	9238	98	9157	9257	65	3748	3948	37,5	3748	4048	27
9172	9222	96	9135	9235	64	3683	3883	37	3749	4049	26,67
9189	9239	96	9158	9258	64	3749	3949	37	3735	4035	26,33
9171	9221	94	9134	9234	63	3682	3882	36,5	3734	4034	26
9190	9240	94	9159	9259	63	3750	3950	36,5	3751	4051	26

Table 2 – Established conserved regions for HIV-1

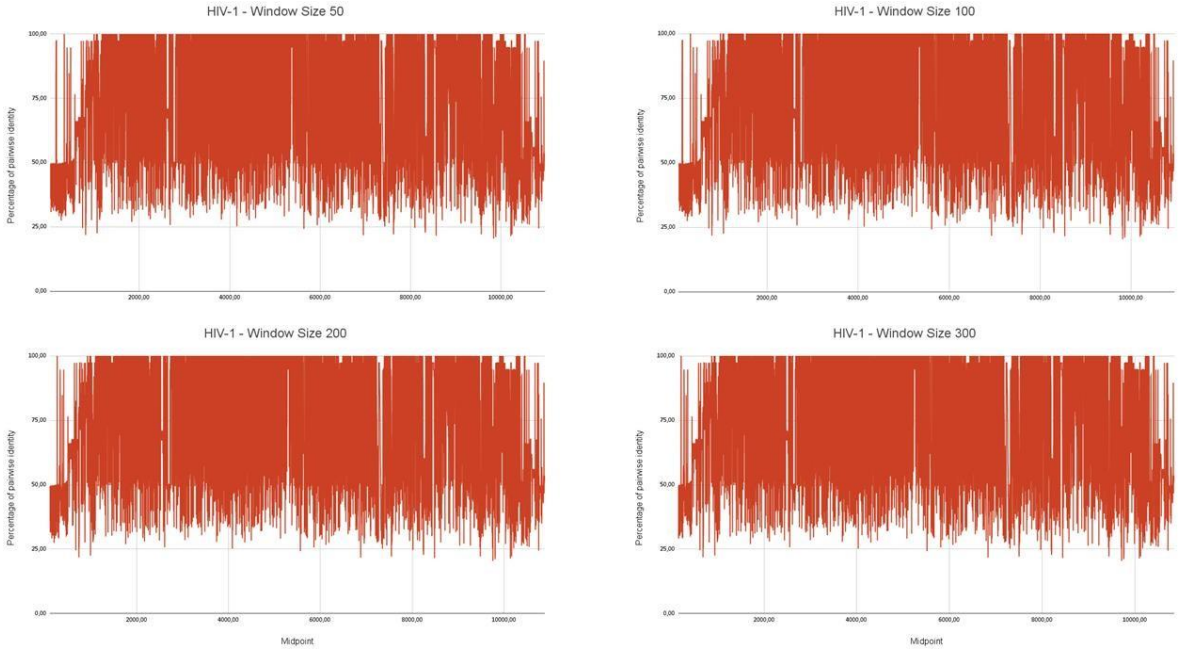


Figure 17 – Percentage of pairwise identity (PPI) for HIV-1 genome

3.1.3. HIV-2

The data for this analysis contained 75 sequences for HIV-2, including the reference sequence. The minimal sequence length was 9,086 and the max sequence length was 10,372. The overall percentage of identical sites of the alignment was 31,6%, and the overall percentage of pairwise identity was 78,4%

For HIV-2 we detected the highest values of percentage of identical sites (PIS) (Figure 18) between 5300 and 5600 nucleotide positions (≈ 70%), and some spikes near 7800 position (≈ 60) and 8800 (≈68) position. The analysis of the genome presented high values of percentage of pairwise identity (Figure 19), with many regions with a value of 100%. The genetic organization of HIV-2 is analogous to that of HIV-1, that is:

5'LTR-*gag*-*pol*-central region-*env*-orf F-3'LTR

The first region with the highest values contains 3 of 5 open reading frames (ORFs), two of them are related to the ORFs of HIV-1, the Q and R regions. This region contains the *vif* gene, responsible for producing a protein critical for infectious virus production in vivo. The spike near 7800bp overlaps with an *env* gene, that encodes the gp160 glycoprotein. This protein is then cleaved into the envelope proteins gp120 and gp41. The former is a viral surface protein that mediates attachment of the virus to target cells. The latter is a transmembrane protein that contains several sites that are required for infection of host cells. The region around 8800bp accommodate the 3'-LTR, and the *nef* gene, a negative regulating factor, responsible for enhancement of infectivity of viral particles, downregulation of CD4 on target cells and with influence on HIV replication (Seitz, 2016). The highest values of PIS and PPI, including the respective window length were annotated (Table 3).

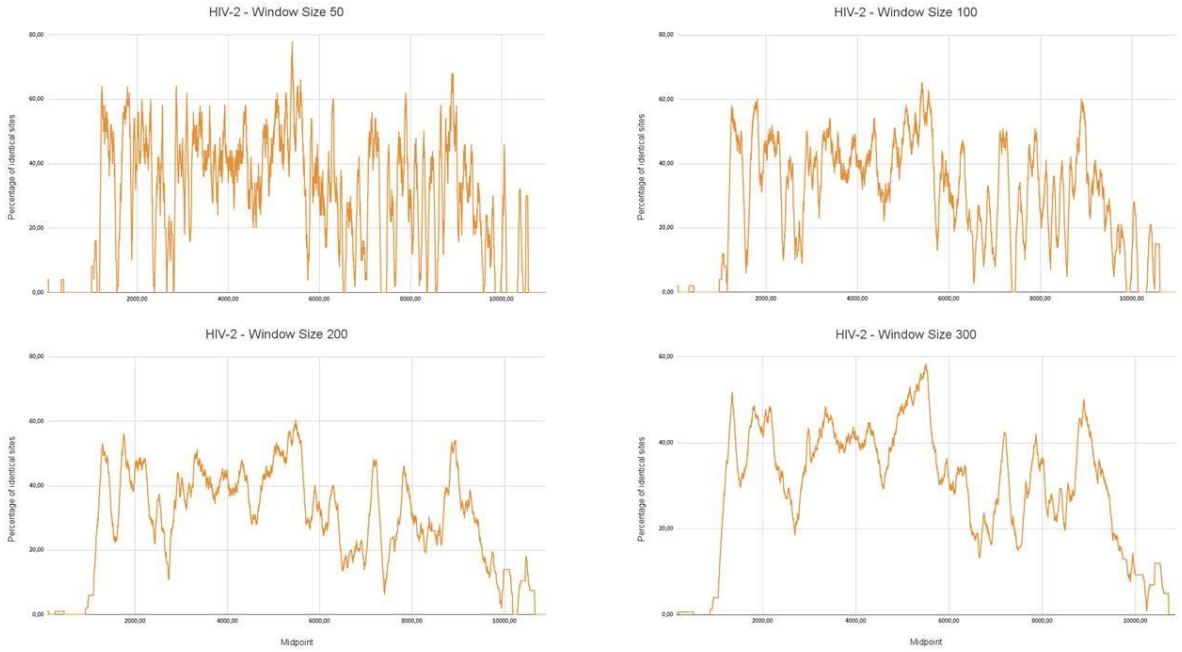


Figure 18 – Percentage of identical sites (PIS) for HIV-2 genome.

Window length											
50			100			200			300		
Start	End	PIS	Start	End	PIS	Start	End	PIS	Start	End	PIS
5379	5429	78	5353	5453	65	5379	5579	60,5	5233	5533	56
5378	5428	76	5358	5458	65	5381	5581	60	5221	5521	55
5370*	5425*	74	5361	5461	65	5382	5582	60	5147	5447	53,33
5361	5411	70	5344	5444	63	5242	5442	57	5013	5313	52,33
8883*	8934*	68	1749	1849	59	5221	5421	56	1201	1501	51,67
5559*	5610*	66	1753	1853	59	5230	5430	56	1205	1505	51
7879	7929	60	8896	8996	58	1197	1397	53	8735	9035	50

Table 3 – Established conserved regions for HIV-2.

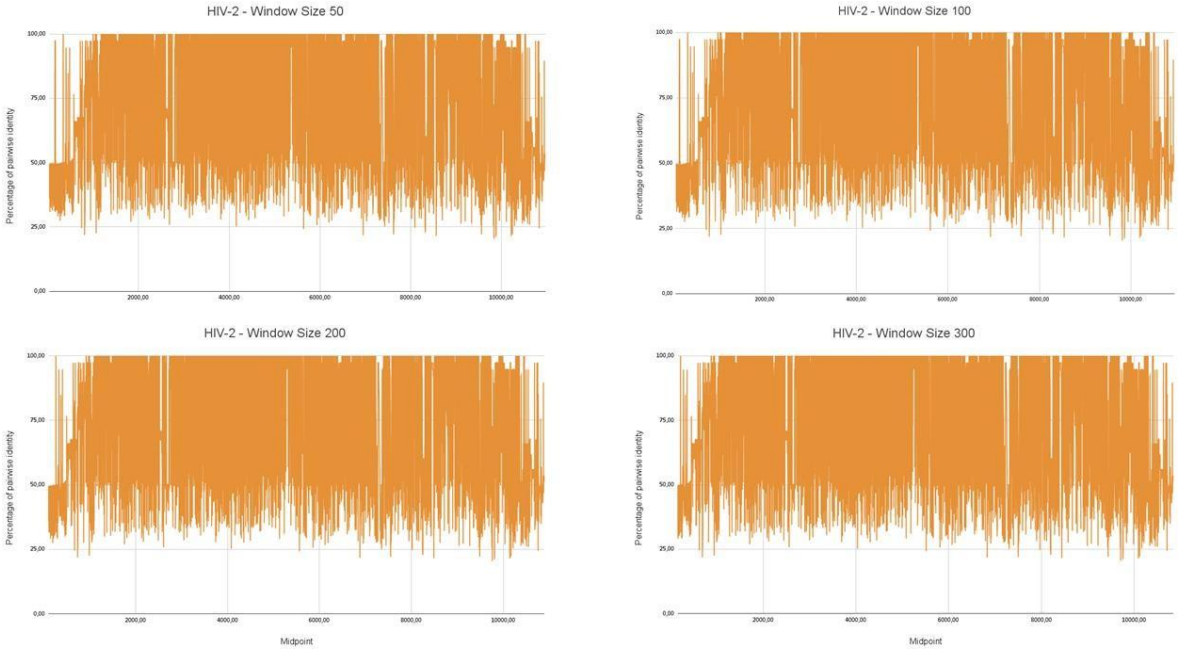


Figure 19 – Percentage of pairwise identity (PPI) for HIV-2 genome.

3.1.4. Ebola

The data for this analysis contained 1,611 sequences for Ebola, including the reference sequence. The minimal sequence length was 18,315 and the max sequence length was 18,959. The overall percentage of identical sites of the alignment was 5,6%, and the overall percentage of pairwise identity was 98,6%.

For the Ebola virus we detected the highest values of percentage of identical sites (PIS) (Figure 20 in the positions ranging from 10600 and 10800 nucleotide positions (≈90%), 14900 and 15100 nucleotide positions (≈ 96%), and some spikes near 16000 and 17000 nucleotide positions (≈81%). In terms of percentage of pairwise identity (PPI) (Figure 21) the Ebola genome presented high value of conservation, with most regions showing a value of 100%.

The first regions contain the *VP24* gene, that is considered to be a secondary matrix protein of Ebola virus and a minor component of mature virions. As a structural protein, *VP24* contributes to virion assembly (Han et al., 2003). The second region comprises a segment of a long gene, the *L* gene. The L protein is thought to be responsible for the enzymatic activities involved in viral RNA replication and transcription (Li et al., 2013). The latest regions are comprised too in this segment, but in the 3'end of the gene. The highest values of PIS and PPI, including the respective window length were annotated (Table 4).

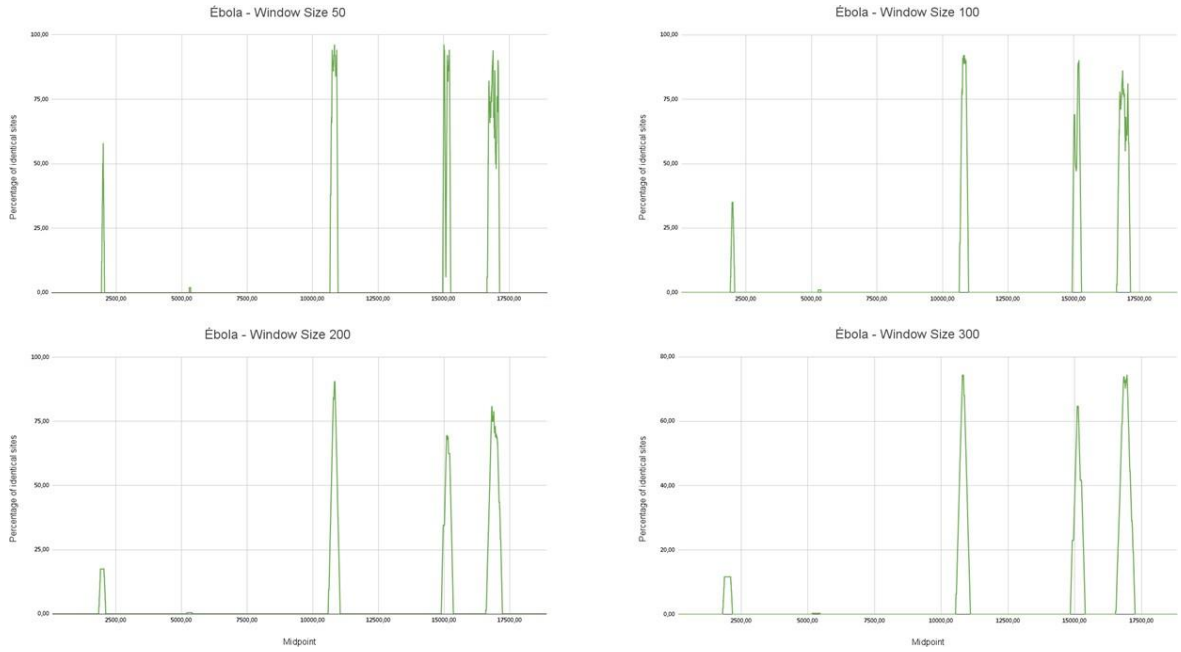


Figure 20 – Percentage of identical sites (PIS) for Ebola virus genome.

Window length											
50			100			200			300		
Start	End	PIS	Start	End	PIS	Start	End	PIS	Start	End	PIS
14987	15037	96	10750	10850	92	10732	10932	90,5	10648	10948	74,33
14988	15038	96	10777	10877	92	10741	10941	90,5	10676	10976	74,33
14989	15039	96	10781	10881	92	10720	10920	90	10641	10941	74
15003	15053	94	10722	10822	91	10744	10944	89,5	10684	10984	74
10718	10768	94	15142	15242	90	16712	16912	81	16807	17107	74
10719	10769	94	15112	15212	89	16703	16903	80	16811	17111	74
10717	10767	92	16814	16914	86	14987	15187	69,5	16798	17098	73,67

Table 4 – Established conserved regions for Ebola

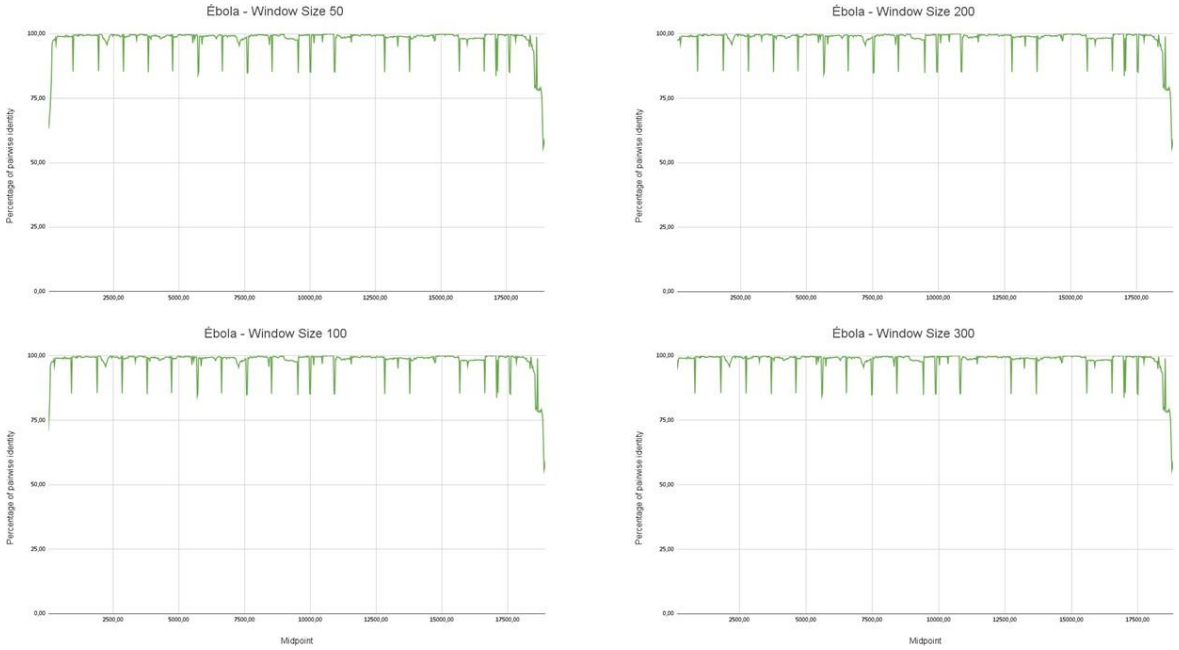


Figure 21 – Percentage of pairwise identity (PPI) for Ebola genome.

3.1.5. Influenza A

The data for this analysis contained 20,601 sequences. The min sequence length was 1,456 and the max sequence length was 2,382. The overall percentage of identical sites was 0,0% and the overall percentage of pairwise identity was 90,0%.
For the influenza A the value for all different windows was 0%. The values for PPI were often high, with various regions with ≈99%, mainly between 100 to 1000 nucleotide positions.

3.1.6. Influenza B

For influenza B the values in the different windows were 0%, on the other end PPI values were constantly with high values (≈96%). The data for this analysis contained 6,287 sequences. The min sequence length was 2,259 and the max sequence length was 2,397. The overall percentage of identical sites was 5,9%, and the overall percentage of pairwise identity was 94,2%. We identified different regions for analysis since the results for PPI were high, ranging from 100 to 500 nucleotides with values ≈ 99%, and from ≈ 2100 to 2180, with values of 100%.

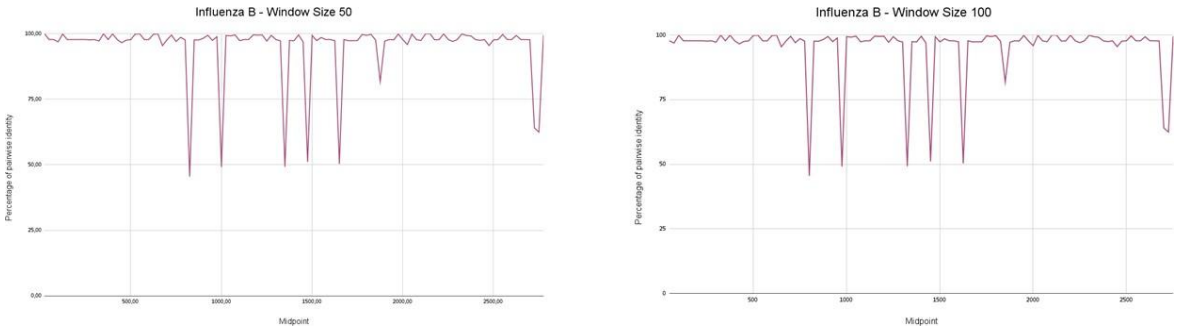


Figure 22 – Percentage of pairwise identity (PPI) for Influenza B genome.

3.1.7. Influenza C

For Influenza C the value in the different windows was 0%, but presented high values of PPI (≈99%). The data for this analysis contained 100 sequences. The min sequence length was 2,325 and the max sequence length was 2,365. The overall percentage of identical sites was 81,9% and the overall percentage of pairwise identity was 97,8%.

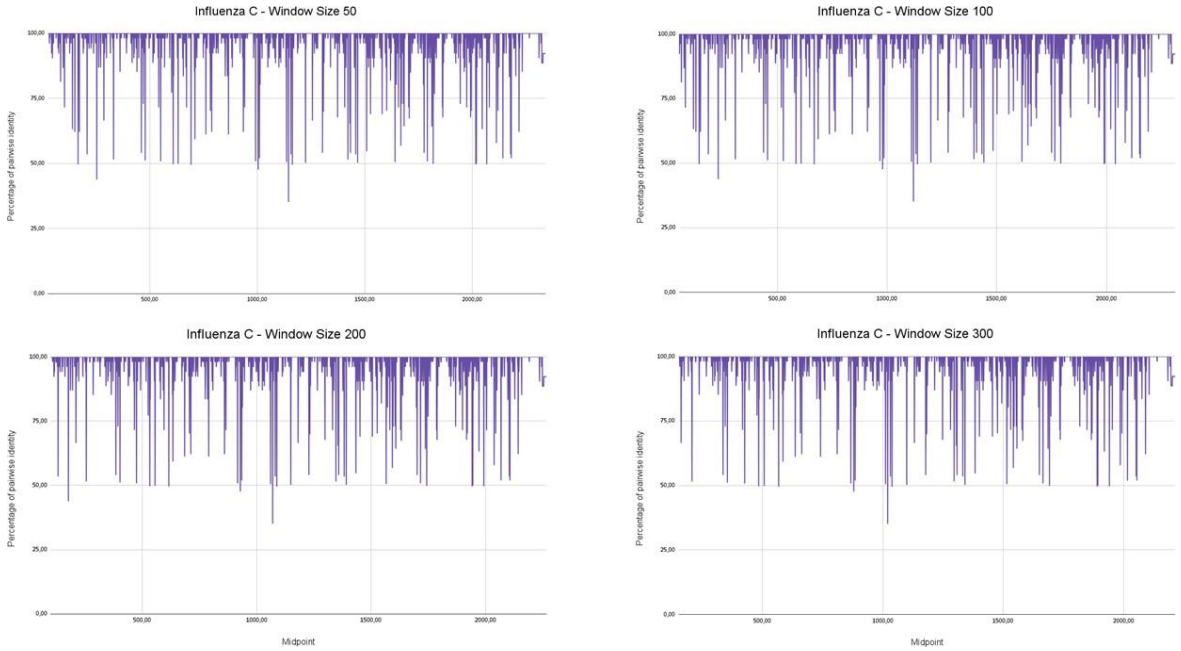


Figure 24 – Percentage of pairwise identity (PPI) for Influenza C.

3.2. Data Input for Primer search

3.2.1. SARS-CoV-2

Start Point	End Point	PIS	PPI	W	Average
28351	28401	100	100	100	100
28751	28801	100	100	100	100
29060	29110	100	100	100	100
27359	27409	100	100	100	100
27780	27830	100	99,23	99,36	99,68
28618	28668	100	99,81	99,84	99,92
29127	29177	100	100	100	100

Table 5 – Data input for SARS-CoV-2.

3.2.2. HIV-1

StartPoint	EndPoint	PIS	PPI	W	Average
1286	1336	6	100	84,33	53
1183	1233	6	100	84,33	53
1638	1688	4	100	84	52
1701	1751	4	100	84	52
3038	3088	4	100	84	52
4226	4276	6	100	84,33	53
7821	7871	12	100	85,33	56

Table 6 – Data-input for HIV-1

For HIV-1 we had to adapt our input due to the fact that the alignment had many gap regions, who were falsely adding conserved regions. Instead, we focused on regions with the highest values of PPI, and with a weighted average value (>=80).

3.2.3. HIV-2

Start Point	End Point	PIS	PPI	W	Average
5379	5429	78	100	96,33	89
5378	5428	76	100	96,0	88
5370	5420	74	100	95,67	87
5361	5411	70	100	95,0	85
8883	8933	68	100	94,67	84
5559	5609	66	100	94,33	83
7879	7929	60	100	93,33	80

Table 7 – Data input for HIV-2.

3.2.4. Ebola

Start Point	End Point	PIS	PPI	W	Average
14987	15037	96	100	99,33	98
14988	15038	96	100	99,33	98
14989	15039	96	100	99,33	98
15003	15053	94	100	99	97
10718	10768	94	100	99	97
10719	10769	94	99,88	98,9	97
10717	10767	92	99,75	98,46	96

Table 8 – Data input for Ebola.

3.2.5. Influenza A

Start Point	End Point	PIS	PPI	W	Average
26	76	0	99,59	82,99	49,795
101	151	0	99,76	83,13	49,88
126	176	0	99,6	83	49,8
201	251	0	99,67	83,05	49,835
426	476	0	99,63	83,025	49,815
626	676	0	99,56	82,96	49,78
701	751	0	99,54	82,95	49,77

Table 9 – Data input for Influenza A.

3.2.6. Influenza B

Start Point	End Point	PIS	PPI	W	Average
205	255	0	97,8	83,33	50
326	376	0	100	83,33	50
2126	2176	0	100	83,33	50
101	151	0	99,94	83,28	49,97
376	426	0	99,97	83,31	49,99
526	576	0	99,94	83,28	49,97
2101	2151	0	99,97	83,31	49,99

Table 10 – Data input for Influenza B.

3.2.7. Influenza C

Start Point	End Point	PIS	PPI	VP	Average
205	255	0	97,8	80,03	48,02
223	273	0	96,04	80,03	48,02
515	565	0	100	83,33	50
622	672	0	100	83,33	50
1117	1167	0	100	83,33	50
1728	1778	0	100	83,33	50
2123	2173	0	100	83,33	50

Table 11 – Data input for Influenza C.

3.3. Primers

3.3.1. SARS-CoV-2

Primer Sequence	Length	G+C Content (40-60%)	Melting Temperature (50-80ºC)	Type
TTGGTCTTACCTCTTTCG	18	47.05	76	Forward
TTGAAGGAGTTCCTTGTT	18	38.88	88	Forward
AGATAAACACGAAAAATC	18	27.77	68	Forward
GCATTGTTAGCAGGATTG	18	44.44	96	Reverse
TACTTCTCGTTGGTTACCT	19	42.10	96	Forward
ATTGGTCTTACCTCTTCCG	19	44.44	92	Forward
GAAGATAAACACGAAAAATC	20	30.0	84	Forward

Table 12 – Generated primers for SARS-CoV-2.

3.3.2. HIV-1

HIV-1 did not present any suitable primers, that fit into the established parameters. All primers generated had high melting temperatures, despite fitting into G+C content. We think that further studies should be made on the alignment data for HIV-1, to better define the conserved regions.

3.3.3. HIV-2

Primer Sequence	Length	G+C Content (40-60%)	Melting Temperature (50-80ºC)	Type
TCCTTCTCCCTTCCACAG	18	55.55	60	Reverse
TCCATCTCCCTTCCACAG	18	55.55	66	Reverse
CCATCCCTGTCTTTATTT	18	38.88	60	Forward
TCCTTCTCCTTTCCACAG	18	50.0	60	Reverse
TTCCATTAAACCCGAACC	18	44.44	76	Reverse
TTCCCCTCCTTGTCCCCT	18	61.11	40	Forward
CTCCTTCTCCCTTCCACAG	19	57.89	64	Reverse
CTCCATCTCCCTTCCACAG	19	57.894	70	Reverse
CTCCTTCTCCTTTCCACAG	19	52.63	64	Reverse
GCTCCTTCCCCTTTCCACAA	20	55.10	76	Reverse
GCTCCTTCCCCTTTCCACAA	20	55.10	76	Reverse
TTTTCCCCTCCTGATCCCCT	20	55.10	56	Forward
TGCTCCTTCCCCCTTCCACAA	21	57.14	80	Reverse
CTTTTCCCCTCCTGATCCCCT	21	57.14	60	Forward
ATTTTCTTCCCCTCCTTGTCCCCT	24	50.0	66	Forward
AATTTTCTTCCCCTCCTTATCCCCT	25	44.0	66	Forward
AATTTTCTTCCCCCCCTTATCCCCT	25	48.0	60	Forward
AATTTTCCTCCCCTCCTTATCCCCT	25	48.0	60	Forward

Table 13 – Generated primers for HIV-2.

3.3.4. Ebola

Primer Sequence	Length	G+C Content (40-60%)	Melting Temperature (50-60ºC)	Type
TGGTTGTGATTGGTAAAG	18	38.88	56	Forward
GGTTGTGATTGGTAAAGT	18	38.88	56	Forward
TGGTTGTGATTGGTGAAG	18	44.44	42	Forward
GGTTGTGATTGGTGAAGT	18	44.44	42	Forward
TTGGTTGTGATTGGTGAA	18	38.88	48	Forward
TTGGTTGTGATTGGTAAAG	19	36.84	64	Forward
TGGTTGTGATTGGTAAAGT	19	36.84	64	Forward
GTTGGTTGTGATTGGTGAA	19	42.11	48	Forward
GTTGGTTGTGATTGGTAAAG	20	40.0	64	Forward
TTGGTTGTGATTGGTGAAGT	20	40.0	54	Forward
TGTTGGTTGTGATTGGTGAA	20	40.0	54	Forward
GTTGGTTGTGATTGGTGAAGT	21	42.86	54	Forward
TTGTTGGTTGTGATTGGTGAA	21	38.10	60	Forward
TTGTTGGTTGTGATTGGTGAAG	22	40.91	60	Forward
TGTTGGTTGTGATTGGTGAAGT	22	40.91	60	Forward
TTGTTGGTTGTGATTGGTGAAGT	23	39.13	66	Forward

Table 14 – Generated primers for Ebola.

3.3.5. Influenza A

Primer Sequence	Length	G+C Content (40-60%)	Melting Temperature (50-80%)	Type
CTTCCGTTCTCTTCTTGG	18	50.0	72	Forward
TACAACCTCTCTCTTGAC	18	44.44	76	Forward
CCTTCCGTTCTCTTCTTG	18	50.0	56	Forward
TTGTCTGCTGTAATTGGG	18	44.44	80	Reverse
GTCCTTCCGTTCTCTTCT	18	50.0	56	Forward
CTCCTTACTTGTTCCTGT	18	44.44	66	Forward
GTCTCTCCTTACTTGTTC	18	44.44	66	Forward
GGTCTCTCCTTACTTGTT	18	44.44	78	Forward
AGTCCCTCTGTCCTCTTCT	19	52.63	80	Forward
TCTCTTTACTCGTTCCTGT	19	42.10	68	Forward
TCTCTTTACTCGTTCCCGT	19	47.36	74	Forward
AGTCCTTCCGTCCTCTTCT	19	52.63	80	Forward
TCTCCTTACTTGTTCCTGT	19	42.10	68	Forward
AGTCCCTCTGTCCTCTTTT	19	47.36	74	Forward
CTTTCTTTACTCGTTCCTG	19	42.10	68	Forward
TCTCCTTACTTGTTCCTGT	19	42.10	68	Forward
AGTCCTTCTGTTCTCTTCT	19	42.10	68	Forward
TTTCCTTACTTGTCCCCGT	19	47.36	74	Forward
AGCCCCTCTGTCCTTTTCT	19	52.63	80	Forward
AGTCCTTCCGTTCTCTTCT	19	47.36	74	Forward
TTTCTTTACTCGTCCCTGT	19	42.10	68	Forward
CTCTCTTTACTCGTTCCTGT	20	45.0	76	Forward
CTCTCTTTACTTGTTCCTGT	20	40.0	70	Forward
TCTCTCCTTACTTGTTCCTGT	21	42.85	78	Forward
CCTCTCTTTACTTGTTCCTGT	21	42.85	78	Forward

Table 15 – Generated primers for Influenza A.

3.3.6. Influenza B

Primer sequence	Lenght	G+C Content (40-60%)	Melting Temperature (50-80ºC)	Type
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTARGGGGTCT	18	41.17	72	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TYTCGTATTAGGGRGTCT	18	43.75	68	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGKATTAGGGGGTCT	18	47.05	76	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTATTAGGGGGTCT	18	44.44	80	Forward
TTTCGTGTTAGGGGGTCT	18	50.0	72	Forward

Table 16 – Generated primers for Influenza B.

3.3.7. Influenza C

Primer sequence	Lenght	G/C Content (40-60%)	Melting Temperature (50-80ºC)	Type
TAAGGGTTTCTTGTGTTG	18	38.88	60	Forward
TTAAGGGTTTCTTGTGTTG	19	36.84	64	Forward
TTTAAGGGTTTCTTGTGTTG	20	35.0	68	Forward

Table 17 – Generated primers for Influenza C.

4. Conclusion

Obtaining good oligonucleotide primers is a challenge, that shall take into account several factors in order to reach maximum efficiency. The progress made in PCR technology and on computational tools allows us to really improve the specificity and efficiency of our protocols. The present work aimed to create a python algorithm, that would generate a list of oligonucleotide primers candidates, for the identification of a set of highly infectious viruses.

The level of conservation for SARS-CoV-2 showed an overall PIS of 93,6% and an overall PPI of 99,95%. The conservation results presented the highest values between all analyzed viruses. That can be explained by the recent evolutionary history of the SARS-CoV-2 and the proofreading activity of SARS. Studies suggest a rate of mutation for SARS-CoV-2 of (1.2 ± 0.5) $x10^{-3}$ mutations per site/per year (Domingo et al., 2021).

For HIV-1, the average values for PIS and PPI were 0,6% and 87,2% respectively. According to (Abram et al., 2010), that HIV-1 high rate of viral replication, the size of the viral population across patients and substantial recombination, contributes to the genetic variation of the genome, what can explain the low percentage of identical sites. HIV-2, however, presented PIS and PPI values of 31,6% and 78,4%. As referred by (Le Hingrat, et al., 2020.) HIV-1 and HIV-2 genomes differ at the nucleotide level by 40% to 60%, so that genomic variation can explain the difference on those results. Also considering that the number of patients across the population, and the fact that HIV-1 is responsible for more than 99% of all HIV infections (Le Hingrat, et al.,2020) the very low value of PIS for HIV-1 is reasonable. A study by (Cuevas et al., 2015) showed an extremely high mutation rate in DNA sequences from peripheral blood mononuclear cells of (4.1 ± 1.7) x 10^{-3} per base per cell, the highest reported for any biological entity.

Ebola also showed low values of PIS with an average of 5,6%, but an average value of PPI of 98,6%. The rate of Ebola is not established, but it is estimated that the rate in humans is ≈4.7 x 10^{-4} substitutions/site/year, across outbreaks since 1976 to 2018 (Whitfield et al., 2020). The studies diverge because, in some outbreaks, the fact that Ebola virus have fatality rates up to 90% (Rodriguez et al., 1999), making it harder to stack mutations, in others suggest that there is evolution, although the mutation rate has been roughly the same over the last outbreaks (Ibrahim, 2014). These results hint that Ebola may have a high degree of genetic variation, most regions on the different windows length showed low values for percentage of identical site, with only 3 clear regions with high conservation measures.

Influenza was the virus that showed the lowest values for PIS, with an average of 0% for the most frequently on the population, Influenza A, 5,9% for Influenza B and 81,9% for Influenza C. Comparing A and B viruses, they are equally prevalent in the population, but is reported that A virus genes evolve two to three times more rapidly than the correlated genes on the B viruses due to a higher mutation rate of A virus of $2.6 \pm 1.2 \times 10^{-3}$ per site/per year. The B variant presented $0.5 \pm 0.4 \times 10^{-3}$ (Nobusawa & Sato, 2006). Influenza C is the less reported type of influenza virus that causes cold symptoms, but it has been reported in pigs, dogs and cattle (ViralZone Swiss Institut of Bioinformatics, 2010). We couldn't find studies on the mutation rate for this genomic variant, possibly because of is rare prevalence on the population and difficulties up to now for detection methods, like RT-PCR, resulted on very few studies on the matter.

Altogether, our data show that the most diverse set of sequences are the Influenza A, HIV-1 and Ebola, which indicates a long evolutionary history of viral adaptation in the populations. Furthermore, intraspecific variation in disease resistance is known to affect pathogen occurrence and may have a larger role in pathogen community structure (Sallinen et al., 2020). For HIV-1, we were not able to detect a sequence block that could be used as primer in PCR reactions. For the remaining viruses, we were able to predict the best primer sequences with reasonable accuracy using genetic conservation data. The primers presented still care laboratory testing to assess their efficiency, as well as further studies on the possibility of secondary structures to avoid internal folding and annealing between primers, which can be done in the future to deepen our knowledge on these results. Nonetheless, we think that the theoretical assumptions, and the optimization of software and algorithms, can lead the way for better and easier diagnoses, giving an advantage for human populations against outbreaks of these organisms. Further research will enhance our understanding and capacity for predicting new pandemics and safety measures.

5. Bibliography

Abram, M. E., Ferris, A. L., Shao, W., Alvord, W. G., & Hughes, S. H. (2010). Nature, Position, and Frequency of Mutations Made in a Single Cycle of HIV-1 Replication. Journal of Virology.

Bachman, J. (2013). Reverse-transcription PCR (RT-PCR). In Methods in Enzymology (1st ed., Vol. 530). Elsevier Inc.

Bartholomeusz, A., & Locarnini, S. (2006). Associated With Antiviral Therapy. Antiviral Therapy.

Beloukas, A., Psarris, A., Giannelou, P., Kostaki, E., Hatzakis, A., & Paraskevis, D. (2016). Molecular epidemiology of HIV-1 infection in Europe: An overview. Infection, Genetics and Evolution.

Blümel, J., Burger, R., Drosten, C., Gröner, A., Gürtler, L., Heiden, M., Hildebrandt, M., Jansen, B., Klamm, H., Montag-Lessing, T., Offergeld, R., Pauli, G., Seitz, R., Schlenkrich, U., Schottstedt, V., Willkommen, H., Von König, C. H. W., & Schweiger, B. (2009). Influenza virus. Transfusion Medicine and Hemotherapy.

Brodin, J., Krishnamoorthy, M., Athreya, G., Fischer, W., Hraber, P., Gleasner, C., Green, L., Korber, B., & Leitner, T. (2013). A multiple-alignment based primer design algorithm for genetically highly variable DNA targets. BMC Bioinformatics.

Carneiro, J., Gomes, C., Couto, C., & Pereira, F. (2020). CoV2ID: Detection and therapeutics oligo database for SARS-CoV-2.

Carneiro, J., & Pereira, F. (2016). EbolaID: An Online Database of Informative Genomic Regions for Ebola Identification and Treatment. PLoS Neglected Tropical Diseases.

Carneiro, J., Resende, A., & Pereira, F. (2017). The HIV oligonucleotide database (HIVoligoDB).

Chang, K. S. (1994). Polymerase chain reaction. Cancer Bulletin.

Chen, D. S., Wu, Y. Q., Zhang, W., Jiang, S. J., & Chen, S. Z. (2016). Horizontal gene transfer events reshape the global landscape of arm race between viruses and homo sapiens. Scientific Reports.

Cleaveland, S., Laurenson, M. K., & Taylor, L. H. (2001). Diseases of humans and their domestic mammals: Pathogen characteristics, host range and the riskof emergence. Philosophical Transactions of the Royal Society B: Biological Sciences.

Clementi, M., Menzo, S., Bagnarelli, P., Manzin, A., Valenza, A., & Varaldo, P. E. (1993). Quantitative PCR and RT-PCR in virology. Genome Research.

Cuevas, J. M., Geller, R., Garijo, R., López-Aldeguer, J., & Sanjuán, R. (2015). Extremely High Mutation Rate of HIV-1 In Vivo.

Dhama, K., Kumar, S., Sharun, K., Pathak, M., & Tiwari, R. (2020). Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ' s public news and information. Elsevier, January.

Dieffenbach, C. W., Lowe, T. M. J., & Dveksler, G. S. (1993). General concepts for PCR primer design. Genome Research.

Domingo, E., Garc, C., & Lobo-vega, R. (2021). Proofreading-Repair Activities in RNA Virus Genetics.

Dziabowska, K., Czaczyk, E., & Nidzworski, D. (2018). Detection methods of human and animal influenza virus—current trends. Biosensors.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research

Green, M. R., & Sambrook, J. (2019). Polymerase chain reaction. Cold Spring Harbor Protocols.

Han, Z., Boshra, H., Sunyer, J. O., Zwiers, S. H., Paragas, J., & Harty, R. N. (2003). Han et al 2003 BIOCHEMICAL AND FUNCTIONAL CHARACTERIZATION OF THE EBOV VP24 PRTEIN IMPLICATION FOR A REOLE IN VIRUS ASSEMBLY AND BUDDING.

Holmes, E. C. (2008). Evolutionary history and phylogeography of human viruses. Annual Review of Microbiology.

Howard, C. R., & Fletcher, N. F. (2012). Emerging virus diseases : can we ever expect the unexpected ?

Ibrahim, H. 2011. F. – faktor yang berhubungan dengan kejadian I. pada anak B. di wilayah P. B. K. B. T. 2011. T. P. P. U. (2014).

Impacts, H.-G., Activity, T., Hingrat, Q. Le, Visseaux, B., Bertine, M., Chauveau, L., Schwartz, O., & Collin, F. (n.d.). crossm Genetic Variability of Long Terminal Repeat Region.

Jacob, S. T., Crozier, I., Fischer, W. A., Hewlett, A., Kraft, C. S., Vega, M. A. de La, Soka, M. J., Wahl, V., Griffiths, A., Bollinger, L., & Kuhn, J. H. (2020). Ebola virus disease. In Nature Reviews Disease Primers.

Jiang, T., & Feng, J. (2013). Algorithms in computational biology. Basics of Bioinformatics: Lecture Notes of the Graduate Summer School on Bioinformatics of China.

Kaushik, A., Tiwari, S., Dev, R., Marty, A., & Nair, M. (2020). Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ' s public news and information.

Kirtipal, N., Bharadwaj, S., & Gu, S. (2020). Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ' s public news and information. Infection, Genetics and Evolution.

Koonin, E. V., Senkevich, T. G., & Dolja, V. V. (2006). The ancient virus world and evolution of cells. Biology Direct.

Kubista, M., Andrade, J. M., Bengtsson, M., Forootan, A., Jonák, J., Lind, K., Sindelka, R., Sjöback, R., Sjögreen, B., Strömbom, L., Ståhlberg, A., & Zoric, N. (2006). The real-time polymerase chain reaction. Molecular Aspects of Medicine.

Li, Z., Xu, J., Chen, Z., Gao, X., Wang, L.-F., Basler, C., Sakamoto, K., & He, B. (2013). The L Gene of J Paramyxovirus Plays a Critical Role in Viral Pathogenesis. Journal of Virology.

Nobusawa, E., & Sato, K. (2006). Comparison of the Mutation Rates of Human Influenza A and B Viruses. Journal of Virology.

Parvez, M. K., & Parveen, S. (2017). Evolution and Emergence of Pathogenic Viruses: Past, Present, and Future. Intervirology.

Prejean, J., Song, R., Hernandez, A., Ziebell, R., Green, T., Walker, F., Lin, L. S., An, Q., Mermin, J., Lansky, A., & Hall, H. I. (2011). Estimated HIV incidence in the United States.

Rodriguez, L. L., De Roo, A., Guimard, Y., Trappier, S. G., Sanchez, A., Bressler, D., Williams, A. J., Rowe, A. K., Bertolli, J., Khan, A. S., Ksiazek, T. G., Peters, C. J., & Nichol, S. T. (1999). Persistence and genetic stability of Ebola virus during the outbreak in Kikwit, Democratic Republic of the Congo, 1995. Journal of Infectious Diseases.

Safiabadi Tali, S. H., LeBlanc, J. J., Sadiq, Z., Oyewunmi, O. D., Camargo, C., Nikpour, B., Armanfard, N., Sagan, S. M., & Jahanshahi-Anbuhi, S. (2021). Tools and techniques for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)/COVID-19 detection. Clinical Microbiology Reviews.

Sallinen, S., Norberg, A., Susi, H., & Laine, A. L. (2020). Intraspecific host variation plays a key role in virus community assembly. Nature Communications.

Seitz, R. (2016). Human Immunodeficiency Virus (HIV). Transfusion Medicine and Hemotherapy.

Shi, M., Lin, X. D., Tian, J. H., Chen, L. J., Chen, X., Li, C. X., Qin, X. C., Li, J., Cao, J. P., Eden, J. S., Buchmann, J., Wang, W., Xu, J., Holmes, E. C., & Zhang, Y. Z. (2016). Redefining the invertebrate RNA virosphere. Nature.

Simon-Loriere, E., & Holmes, E. C. (2011). Why do RNA viruses recombine? Nature Reviews Microbiology.

Te Velthuis, A. J. W., & Fodor, E. (2016). Influenza virus RNA polymerase: Insights into the mechanisms of viral RNA synthesis. Nature Reviews Microbiology.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. Nucleic Acids Research.

Valones, M. A. A., Guimarães, R. L., Brandão, L. A. C., De Souza, P. R. E., De Albuquerque Tavares Carvalho, A., & Crovela, S. (2009). Principles and applications of polymerase chain reaction in medical diagnostic fields: A review. Brazilian Journal of Microbiology.

ViralZone Swiss Institut of Bioinformatics. (2010). Influenza C virus genome.

Watzinger, F., Ebner, K., & Lion, T. (2006a). Detection and monitoring of virus infections by real-time PCR.

Watzinger, F., Ebner, K., & Lion, T. (2006b). Detection and monitoring of virus infections by real-time PCR. Molecular Aspects of Medicine.

Whitfield, Z. J., Prasad, A. N., Ronk, A. J., Kuzmin, I. V., Ilinykh, P. A., Andino, R., & Bukreyev, A. (2020). Species-Specific Evolution of Ebola Virus during Replication in Human and Bat Cells. Cell Reports.

Wu, D., Wu, T., Liu, Q., & Yang, Z. (2020). International Journal of Infectious Diseases The SARS-CoV-2 outbreak : What we know.

Yildirim, D., Sagdic, D. O., Seflek, B., Cimentepe, M., Bayram, I., & Yarkin, F. (2017). Influenza Virüs Enfeksiyonlannin Moleküler ve Immün Floresan Yöntemlerle Saptanmasi. Mikrobiyoloji Bulteni.

Zheng, L., Wayper, P. J., Gibbs, A. J., Fourment, M., Rodoni, B. C., & Gibbs, M. J. (2008). Accumulating Variation at Conserved Sites in Potyvirus Genomes Is Driven by Species Discovery and Affects DegeneratePrimer Design.