FACULTY OF
BIOSCIENCE ENGINEERING

COMPUTATIONAL EXPLORATION TOWARDS UNIVERSAL VACCINES FOR INFLUENZA AND CORONAVIRUSES

SUBTITLE???

word count: ???

Menno Van Damme
Student ID: 01709485

Promotor: Prof. Dr. Ir. Wim Van Criekinge
Tutor: Dr. Christian Rausch

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of master in Bioscience Engineering.
Academic year: 2021 - 2022

GHENT
UNIVERSITY

Gent, FILL IN THE DATE

The promotor,	The author,
Prof. Dr. Ir. Wim Van Criekinge	Menno Van Damme

ACKNOWLEDGEMENTS

Some of the computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI.

CONTENTS

ABBREVIATIONS

CDS Coding sequence

COVID-19 Coronavirus disease 2019

cRNA Complementary RNA

DdRp DNA-dependant DNA polymerase

DdRp DNA-dependant RNA polymerase

DNA Deoxyribonucleic Acid

dsDNA Double-stranded DNA

dsRNA Double-stranded RNA

E Envelope protein

ER Endoplasmatic reticulum

HA Hemagglutinin

HPC Central High Performance Computing Infrastructure of Ghent University

IAV Influenza A virus

kb Kilobases

M Membrane protein

M1 Matrix protein 1

M2 Matrix protein 2

MERS-CoV Middle East respiratory syndrome-related coronavirus

MHC Major histocompatibility complex

mRNA Messenger RNA

MSA Multiple sequence alignment

N Nucleocapsid protein

NA Neuraminidase

NCBI National Center for Biotechnological Information

NEP Nuclear export protein

NP Nucleoprotein

NS1 Non-structural protein 1

NS2 Non-structural protein 2

NSP Non-structural protein

ORF Open reading frame

P1 Acidic polymerase protein

PB1 Basic polymerase protein 1

PB2 Basic polymerase protein 2

RdRp RNA-dependant DNA polymerase

RdRp RNA-dependant RNA polymerase

RNA Ribonucleic acid

RNA polIII RNA polymerase III

S Spike protein

SA Sialic acid

SARS Severe acute respiratory syndrome

SARS-CoV *Severe acute respiratory syndrome coronavirus*

SARS-CoV-1 severe acute respiratory syndrome coronavirus 1

SARS-CoV-2 Severe acute respiratory syndrome coronavirus 2

ssDNA Single-stranded DNA

ssRNA Single-stranded RNA

UTR Untranslated region

vRNA Viral RNA

vRNP Viral ribonucleoprotein

SAMENVATTING

nederlandse samenvatting

SUMMARY

english summary

1. INTRODUCTION

1.1 General introduction to viruses

Viruses are the most abundant biological entities on Earth. They are present everywhere and are able to infect all cellular life (Harris and Hill, 2021).

Viruses exhibit high copy error rates, have large population sizes and have short generation times. Because of this they are experts at quickly adapting to new environments (Manrubia and Lázaro, 2006). These traits allow viruses to quickly evolve and generate diversity which gives them the ability to evade our immune systems, existing vaccines and therapies.

1.1.1 Viral structure

A single viral particle outside of it's host cell is called a virion. It consist of two or three parts: one or more nucleic acid molecules, a protein coat called a capsid and in some cases a lipid bilayer which envelops the capsid (the viral envelope) (Manrubia and Lázaro, 2006).

The nucleic acid molecule is the viral genome and carries the genetic information of the virus. The genome consist of RNA (ribonucleic acid) or DNA (deoxyribonucleic acid) and these can be either single-stranded (ssRNA and ssDNA) or double-stranded (dsRNA and dsDNA). Single-stranded RNA can also be divided into positive-sense or negative-sense, abbreviated as ssRNA(+) and ssRNA(-), which relates to in which direction the RNA can be used to produce protein (Manrubia and Lázaro, 2006).

The viral genome codes for two types of proteins: structural ones and nonstructural ones. Structural proteins are the ones that make up the extracellular virion. Nonstructural proteins function during the viral replication cycle but are not included in the final virion. These are responsible for a wide variety of tasks inside the host cell such as replicating the viral genome, regulating viral assembly and modulating the host antiviral response (Payne, 2017).

capsid shape? enveloped and non-enveloped?

1.1.2 Viral replication cycle

In order to replicate a virus needs to infect a host cell. It can then use the cellular machinery of the host cell to copy it’s genome and produce more viral proteins which then assemble into new virions. There are large differences in replication cycles between different viruses but they all include the following necessary steps (Manrubia and Lázaro, 2006; Payne, 2017).

—-picture of cycle, biorender?—-

1. Attachment
Viruses are usually specific towards certain hosts and host cells. This is called host tropism and cell tropism. This tropism is caused by a specific binding between cell surface receptors and viral capsid or envelope proteins, a process called attachment.

2. Entry into the cell
Viruses have three ways of entering a cell: via genome injection, via endocytosis or via membrane fusion.

During genome injection the capsid or envelope proteins of the virus create a pore in the plasma membrane of the host cell. The viral genome is then injected into the cytoplasm leaving the rest of the virion behind. This method of entry is most common in bacteriophages.

Viruses can also enter using the innate endocytosis process of the cell. Here, the entire virion is taken up by the cell into a vesicle. The virus is then able to escape the vesicle into the cytoplasm.

Membrane fusion is limited to enveloped viruses. After cell attachment the viral envelope merges with the cell membrane and the capsid is released into the cytoplasm. The merging of the membranes is facilitated by the glycoproteins in the viral envelope.

3. Uncoating
Depending on the method of entry, uncoating may or may not be necessary.

It consists of the disassembly of the proteins (and possible envelope) that surround the genome so that it can be copied in the next step. Uncoating can happen in the cytosol or in the nucleus.

4. Replication and expression of the viral genome

In order to produce new virus particles the genome has to be replicated and it's genes expressed into proteins. The way this can be done varies a lot between the different types of viruses. On overview of these strategies is given in figure 1.1.

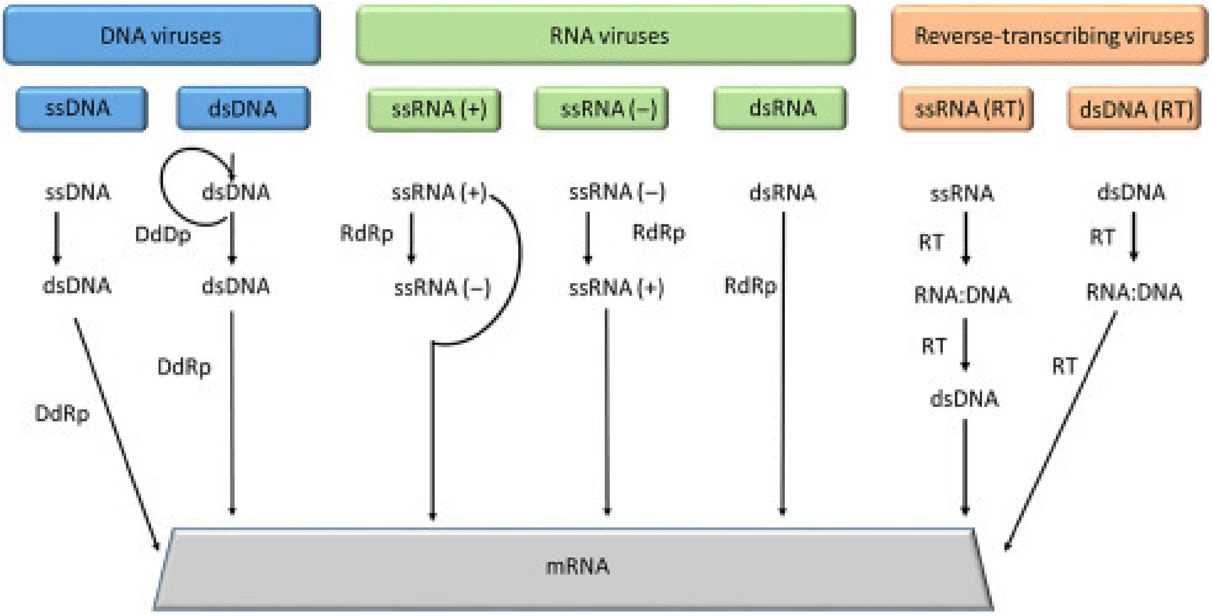


Figure 1.1: Summary of replication and transcription modes of different classes of viruses. Retrieved from Rampersad and Tennant (2018).

DNA viruses rely on DNA-dependent DNA polymerase (DdDp) to replicate their genome. Some viruses use their host's enzymes to replicate while others code for their own version of the protein. RNA viruses copy their genomes using RNA-dependant RNA polymerase (RdRp), also called RNA replicase, which is not usually present in the host cell so their genome contains the genes to make the enzyme.

In order to produce viral proteins the coding parts of the viral genomes has to be transcribed into messenger RNA (mRNA) which acts as a template for the production of protein during translation. In DNA viruses this is done by the host cell's DNA-dependant RNA polymerase (DdRp), called RNA polymerase III (RNA polIII). In RNA viruses this is done by the same RdRp that replicates their genome except in this case only a coding part of the genome is replicated. ssRNA(+) can directly be used by the host cell as mRNA, ssRNA(-) first has to be transcribed into ssRNA(+) which can then be used as mRNA.

Another category of viruses are the reverse-transcribing viruses. these include the retroviruses and hepadnaviruses. They are named this way because during their replication cycle they convert RNA into DNA. Retroviruses are ssRNA(+) viruses of the family *Retroviridae*. They use a RNA-dependant DNA polymerase (RdDp), also called reverse-transcriptase (RT), to convert their RNA genome into dsDNA. This dsDNA then integrates into the genome of the host cell and can then be transcribed back into RNA by RNA polIII. This RNA can then function as mRNA or genomic RNA. Hepadnaviruses are dsDNA viruses of the family *Hepadnaviridae* that use a polymerase which has both DdRp and RdDp activity. This enzyme converts their dsDNA genome into ssRNA. Once the RNA is integrated into a new capsid it gets reverse-transcribed back into dsDNA.

5. Assembly

Once the concentrations of viral genomes and structural proteins are sufficiently high inside the host cell the viruses start to assemble into particles. The assembly sometimes takes place in specific cell regions called viral factories. The cytoskeleton of the cell and various cell organelles are usually heavily involved. RNA viruses often package their genome, this is done to discriminate between mRNA and genomic RNA. This packaging is facilitated by nucleocapsid proteins which recognise packaging signals at the ends of the viral genome.

6. Release

Different viruses employ different strategies to leave the cell. Enveloped viruses are usually released by budding from the plasma membrane or via exocytosis. In budding, the viral nucleocapsid interacts with a region of the cell plasma membrane where glycosylated viral envelope proteins have been inserted. This specific interactions causes the nucleocapsid to be released from the cell surround by a part of the plasma membrane, which forms the envelope. Exocytosis is similar to budding but does not occur at the plasma membrane. This time budding occurs at endosomal and nuclear membranes from which the viruses can enter vesicles. These vesicles travel to the plasma membrane and fuse with it, releasing the virions. Non-enveloped viruses usually escape via lysis of the infected cell, which is caused by the viral infection itself.

1.1.3 RNA viruses

The two viruses studied in this thesis are the influenza A virus and the severe acute respiratory syndrome coronavirus 2, both of which are RNA viruses. RNA viruses have some specifically intertesting features which influence their evolution: they produce their own replicative machinery, they have very high mutation rates, their genome is highly structured and this structure is conserved.

Evolutionary mechanisms in viral genomes

RNA structure

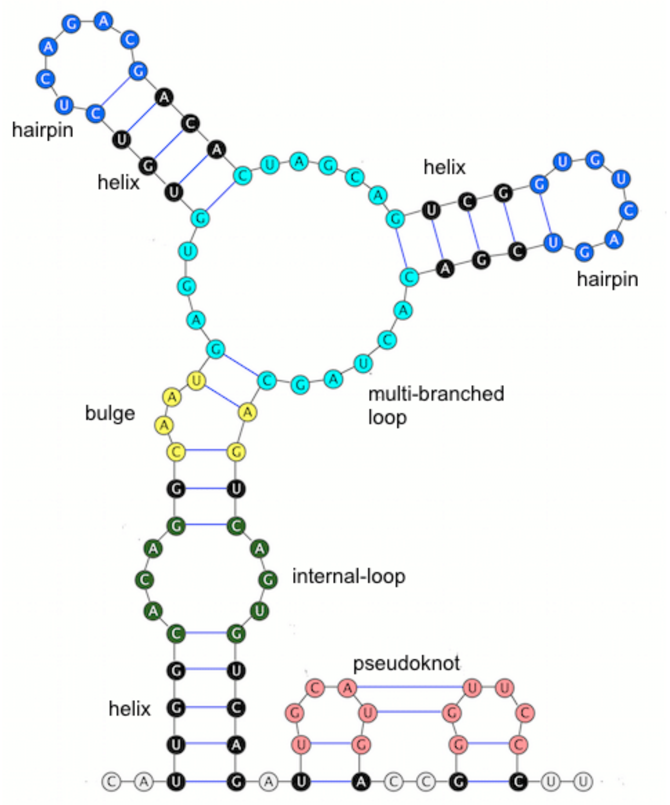


Figure 1.2: The different types of RNA structure elements. Retrieved from Mamuye et al. (2016).

1.2 Influenza A virus

The influenza A virus (IAV) is the only species of the genus *Alphainfluenzavirus* in the family of *Orthomyxoviridae* which also includes the influenza B, C and D virus types. The A, B and C types are all able to cause the influenza disease in humans, more commonly called the ’flu’. The IAV is the dominant causative agent for the disease in humans and is able to infect a wide variety of other mammalian and avian species (Suarez, 2016).

The IAV type is further divided into different subtypes depending on the antigenic variation in the hemagglutinin (HA) and neuraminidase (NA) glycoproteins on the surface of the virus. There are 16 subtypes of HA and 9 subtypes of NA. In the naming of the virus subtypes these are abbreviated to ’H’ and ’N’ respectively. The only subtypes that are known to infect in humans are H1N1, H2N2, H3N2, H5N1, H7N7 and H9N2 (Cheung and Poon, 2007).

1.2.1 Virion and genome structure

All *Orthomyxoviridae* are enveloped and have a segmented ssRNA(-) genome (Cheung and Poon, 2007). The IAV genome exists out of eight segments and codes for 10 to 14 proteins, depending on the subtype (Eisfeld et al., 2015). The segment lengths range from approximately 890 to 2341 nucleotides and the total length of the genome amounts to approximately 13.5 kilobases (kb) (Ghedin et al., 2005).

The segments are numbered according to decreasing length and each encodes for different proteins (Eisfeld et al., 2015; Cheung and Poon, 2007), these are listed below and illustrated on figure 1.3:

- Segment 1: basic polymerase protein 2 (PB2)
- Segment 2: basic polymerase protein 1 (PB1)
- Segment 3: acidic polymerase protein (PA)
- Segment 4: hemagglutinin (HA)
- Segment 5: nucleoprotein (NP)

- Segment 6: neuraminidase (NA)
- Segment 7: matrix protein 1 (M1) & matrix protein 2 (M2)
- Segment 8: nonstructural protein 1 (NS1) & nonstructural protein 2 (NS2) also known as nuclear export protein (NEP)

M2 and NEP are produced by alternative splicing of the M and NS mRNA respectively. Some IAV subtypes also code for other accessory proteins such as PB1-F2, PB1-N40, PA-X, PA-N182, PA-N155, PB2-S1, M42, and NS3 (Dou et al., 2018; Eisfeld et al., 2015). These are produced by the transcription of alternative reading frames or via alternative splicing. They provide additional functions such as suppressing host immunity, increasing viral protein production and promoting cell death and thus influence the virulence of the virus (Khaperskyy et al., 2016; Varga and Palese, 2011).

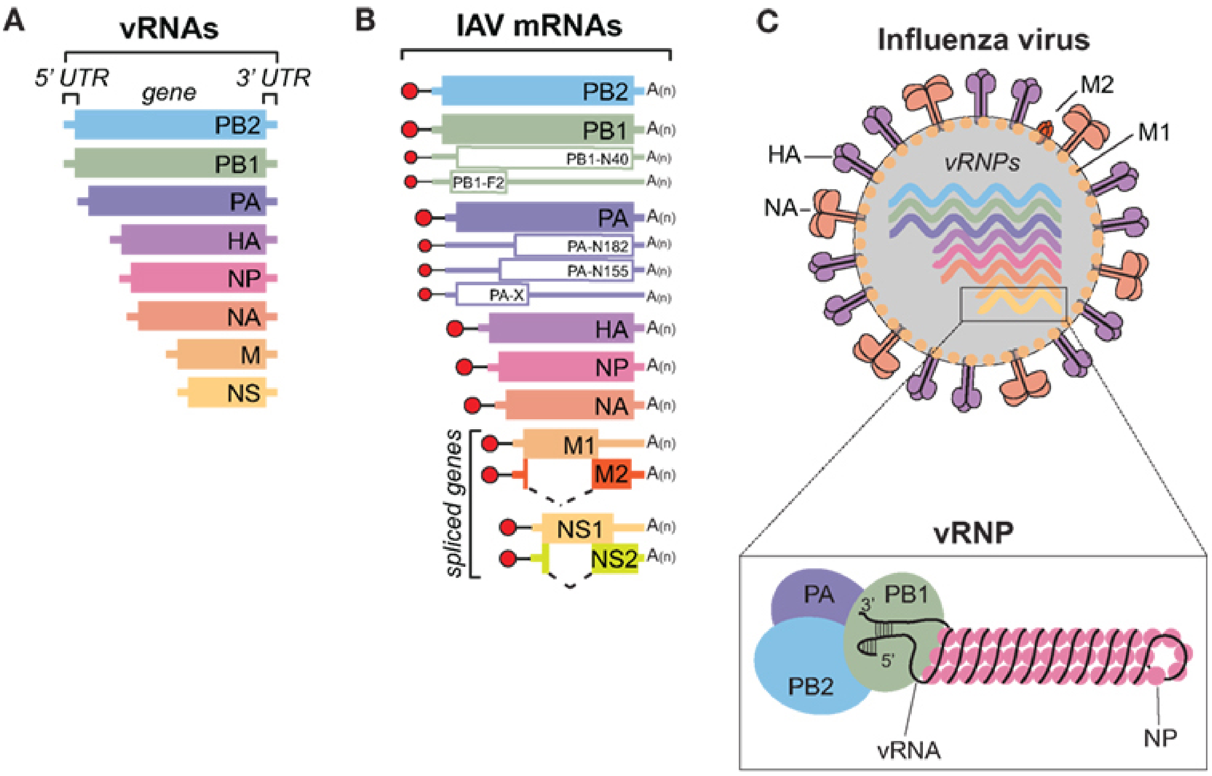


Figure 1.3: (A) Schematic of the eight viral RNA (vRNA) genome segments of the IAV. The 5' and 3' untranslated region (UTR) are represented with a line and contain the viral promoters. The box corresponds to the translated region. (B) Schematic of the viral mRNAs that are transcribed from the vRNA templates. Boxes are the protein coding regions. Dashed lines show the alternative splicing of the M and NS mRNA. The red circles with black lines represent the 5' cap and the host-derived 10-13 nucleotide regions that are obtained from host mRNA by the viral polymerase. The 3' poly-A tail is denoted as A(n). The white colored boxes represent some smaller mRNA transcripts that encode accessory proteins. (C) Schematic of an the IAV virion. The viral membrane proteins HA, NA, and M2 are shown, along with the eight viral ribonucleoproteins (vRNPs) and the matrix protein M1 that supports the viral envelope. A single vRNA gene segment is shown wrapped around multiple nucleoprotein (NP) copies. The conserved promoter regions in the 5' and 3' UTRs base-pair and are bound by a single RdRp complex consisting of PB1, PB2, and PA. Adapted from Dou et al. (2018).

As can be seen on figure 1.3, the eight genome segments are packaged as viral ribonucleoprotein (vRNP) complexes. These consist of the genomic viral RNA (vRNA) wrapped around a a helical structure made up of multiple NPs and a single RdRp complex consisting of its three subunits PB1, PB2 and PA (Eisfeld et al., 2015; Dou et al., 2018). The vRNPs are bound to the viral envelope by the M1 protein, which associates with both and gives structure to the virion (Calder et al., 2010).

The viral envelope is embedded with the HA and NA glycoproteins and M2 (Cheung and Poon, 2007). HA is responsible for binding to cellular receptors and facilitating uptake of the virus by the cell and fusion with the endosomal membrane (Cheung and Poon, 2007; Byrd-Leotis et al., 2017). The role of NA is not yet completely elucidated but it has been shown that it is an essential protein in the virulence and the morphology of the IAV (Jin et al., 1997; Mitnaul et al., 1996). NA also aides HA in recognition the cell receptor (Dou et al., 2018). The M2 protein is a proton selective ion channel and regulates the pH inside the virion which is necessary for the disassociation of the M1 protein from the vRNPs. M2 also functions in the budding of the IAV from the cell (Lamb et al., 1985).

The NS1 protein is the only nonstructural protein of the IAV. In early studies it was believed that NEP was also nonstructural but more recent research has shown that it is incorporated into virions in low amounts (Cheung and Poon, 2007). NS1 is mainly localized in the nucleus where it binds to RNA molecules. By doing so the protein stimulates the production of viral proteins and represses the production of cellular proteins. This way the protein also defends the virus against the cell's antiviral response by repressing the production and activity of interferons which inhibit viral reproduction (Cheung and Poon, 2007; Krug, 2015). The NEP is responsible for the export of vRNPs from the nucleus together with M1 (Neumann et al., 2000). It has also been shown to be involved in regulating viral RNA transcription and translation and in the budding of the virus (Paterson and Fodor, 2012).

1.2.2 Replication cycle

1. Attachment
When the IAV reaches a host cell its HA protein binds to terminal sialic acid (SA) monosaccharides in glycoconjugates on the the cell surface (Dou et al., 2018). NA aides in searching for the correct receptor by removing unproductive SA interactions using its sialidase activity (Gamblin and Skehel, 2010). The binding of HA to the receptor is depends on the way that the SA is linked to the neighbouring monosaccharides in the glycan chain. This SA-linkage is host specific and thus HA partly determines the host tropism of the IAV (Suarez, 2016). Avian IAVs tend to prefer α-2,3-linked SA while human IAVs prefer α-2,6 linkage (Dou et al., 2018).

2. Entry into the cell

The binding of HA to its receptor triggers the entry of the virus into the cell by receptor-mediated endocytosis. The virion ends up on the inside of an endosome, bound to its membrane by the bond between HA and the glycans (Suarez, 2016).

3. Uncoating

The endosome acidifies which activates the M2 ion channel and leads to the uptake of protons into the virion. This causes a conformational change in HA which triggers fusion of the viral envelope and the endosomal membrane (Gamblin and Skehel, 2010). The acidification of the virion also leads to the dissociation of M1 from the vRNPs and their subsequent release into the cytoplasm of the cell (Dou et al., 2018).

4. Replication and expression of the viral genome

The NP carries nuclear localization signals which guides the vRNPs into the nucleus via the importin-α-importin-β nuclear import pathway (Melén et al., 2003).

In the nucleus vRNA is first transcribed into mRNA (Dou et al., 2018). In order to produce viable mRNA transcripts they have to be capped at the 5' end. To obtain the 5' cap the viral polymerase performs a process called 'cap snatching'. Here the PB2 and PA subunit bind and cut a part of the 5' untranslated region (UTR) of the host's pre-mRNA which functions as a capped primer. PB1 then binds this complex forming a fully functional RdRp which then elongates the primer into viral mRNA using the vRNA as a template (De Vlugt et al., 2018).
—– poly A tail —-
The produced mRNA is then exported from the nucleus into the cytoplasm where it is translated by ribosomes into protein (Dou et al., 2018).

At a later stage in the infection the vRNA is also replicated. the negative-sense vRNA is transcribed into positive-sense complementary RNA (cRNA) by the RdRp. This cRNA forms the template for further replication of the vRNA (Dou et al., 2018).

5. Assembly

The newly produced PA, PB1, PB2 and NP are imported into the nucleus and assemble into new vRNPs together with the replicated vRNA (Eisfeld et al., 2015). M1 and NEP are also transported to the nucleus to aide in the export of the vRNPs to the cytosol. Once exported, the vRNPs then associate with the Rab11 protein which transports the vRNPs to the apical cell membrane (Dou et al., 2018).

The envelope proteins HA, NA and M2 are produced by endoplasmatic reticulum (ER) associated ribosomes. After synthesis and folding in the ER the proteins are transported to the Golgi apparatus where post-translation modification takes place. Afterwards they are transported in vesicles towards the apical cell membrane and integrated therein (Dou et al., 2018).

All the structural proteins and vRNPs associate in or around lipid rafts in the apical membrane (Rossman et al., 2010; Leser and Lamb, 2005). The M1 protein binds to both the membrane and cytosolic proteins and thus forms a bridge between the internal and external components of the virion (Rossman and Lamb, 2011).

6. Release
HA, NA, M1 and M2 alter the curvature of the membrane which leads to budding of the virus from the cell (Rossman and Lamb, 2011). After budding HA is still bound to SA on the surface of the cell. These interactions are broken Using the sialidase property of NA and the progeny virus is released from the cell (Dou et al., 2018).

1.2.3 Evolution

1.2.4 Vaccines

1.3 Severe acute respiratory syndrome coronavirus 2

Severe acute respiratory syndrome coronavirus 1 (SARS-CoV-1) and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) are two strains of the species *Severe acute respiratory syndrome–related coronavirus* (SARS-CoV). SARS-CoV belongs to the family of *Coronaviridae*, the subfamily of *Orthocoronavirinae* and the genus *Betacoronavirus* together with the well-known *Middle East respiratory syndrome-related coronavirus* (MERS-CoV) and several other species. Both SARS-CoV strains are responsible for respiratory diseases in humans. SARS-CoV-1 causes the severe acute respiratory syndrome (SARS) and SARS-CoV-2 causes the coronavirus disease 2019 (COVID-19) (of the International Committee on Taxonomy of Viruses, 2020).

—— Pangolin lineages, WHO classification, Nextstrain clades, GISAID clades

1.3.1 Virion and genome structure

All Coronaviridae have a ssRNA(+) genome with a size ranging from approximately 27 to 32 kb (Brian and Baric, 2005). The genome of SARS-CoV-2 has a an approximate length of 29.9 kb (Lu et al., 2020) and codes for four structural proteins, sixteen non-structural proteins (NSPs) and nine accessory proteins (Yadav et al., 2021; Bai et al., 2021). A schematic overview of the genome is shown in figure 1.4. The genome contains two large open reading frames (ORFs) at the 5' end: ORF1a and ORF1b. Translation of ORF1a produces a polyprotein called pp1a. ORF1b is only translated when a -1 ribosomal frameshift occurs during the translation of ORF1a which allows the ribosome to continue beyond the stop codon. When this occurs this results in the production of a second polyprotein pp1ab. pp1a and pp1ab are co- and post-translationally processed via autoproteolytic cleavage into the sixteen NSPs: NSP1 up to NSP16 (Sola et al., 2015).

The four structural proteins are the spike protein (S), envelope protein (E), membrane protein (M) and nucleocapsid protein (N) which make up the virion (Wang et al., 2020). A diagram of the virion structure can be seen in figure 1.5.

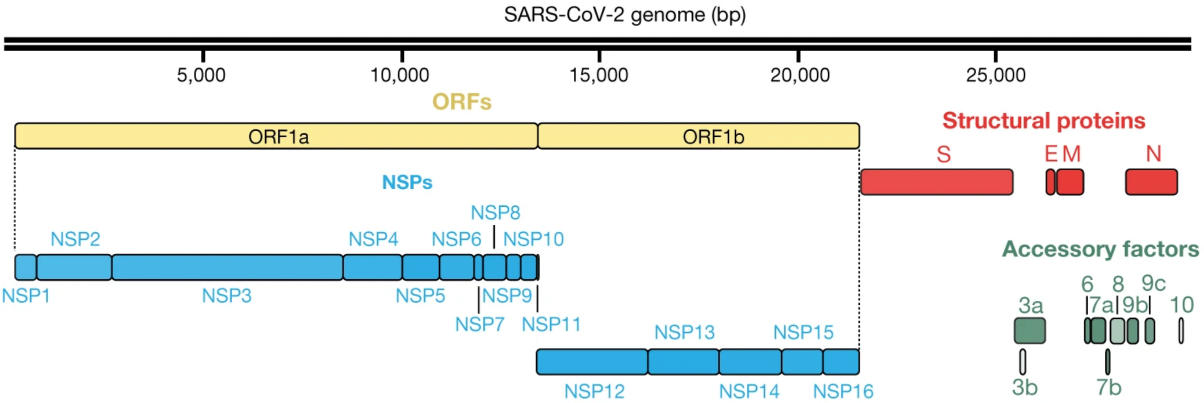


Figure 1.4: SARS-CoV-2 genome. Adapted from Gordon et al. (2020).

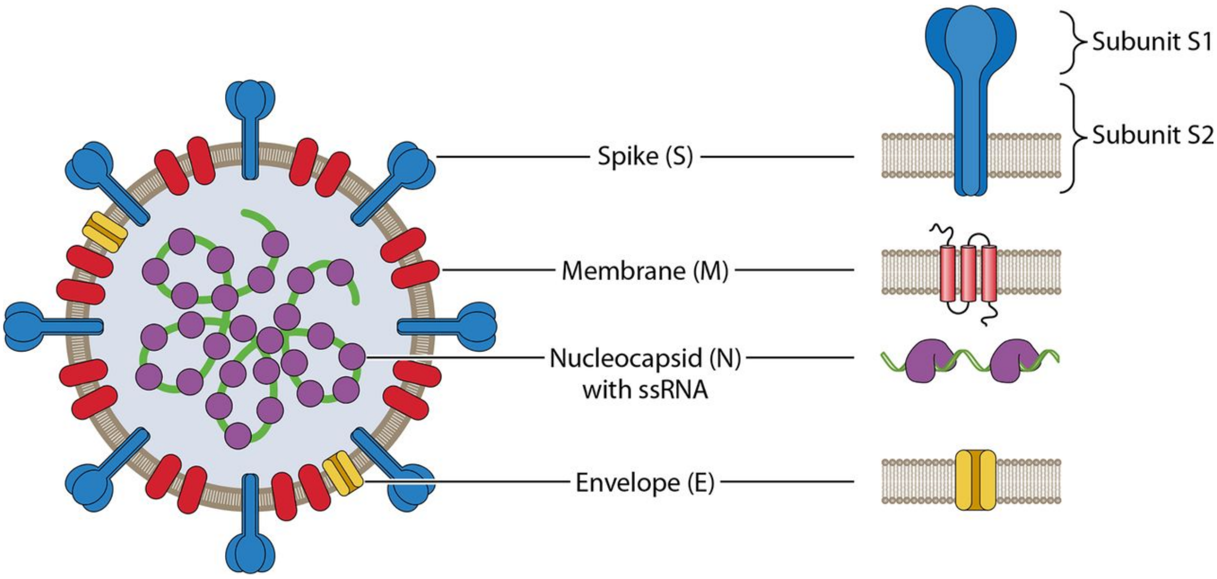


Figure 1.5: SARS-CoV-2 structure. Retrieved from Synowiec et al. (2021).

1.3.2 Replication cycle

1. Attachment

2. Entry into the cell

3. Uncoating

4. Replication and expression of the viral genome

5. Assembly

6. Release

1.3.3 Evolution

1.3.4 Vaccines

1.4 Universal vaccines

2. GOALS AND OBJECTIVES

The main objective of this thesis is to identify highly conserved, protein coding regions in the RNA genome of the IAV and SARS-CoV-2. Due to the conserved nature of these regions across all variants of the virus they would have the potential to be used as a universal vaccine. Since the RNA secondary structure of ssRNA viruses is typically a conserved feature we also wanted to investigate whether there was a correlation between the secondary structure elements and how conserved these regions were.

We also aim to investigate the relationship between the secondary structure of the RNA genome and its conservation for both viruses.

RNA sequences for both viruses were retrieved from the National Center for Biotechnology Information (NCBI) (Sayers et al., 2010). Only sequences that had complete coverage of the coding regions were used in the analysis. Additionally, only sequences that were collected from human hosts and had a complete collection date were retained. Sequences that contained ambiguous characters as well as duplicate sequences were excluded. In the case of duplicate sequences the sequence with the earliest collection date was kept.

Thereafter a multiple sequence alignment (MSA) was created using the MAFFT software (Katoh et al., 2002). The MSA was trimmed by removing rare insertions as to have approximately the same number of positions as the average length of the unaligned genome sequences. From the MSA the conservation of each position in the genome was scored using two metrics, the Shannon entropy and the mutability. For both metrics the total conservation score was calculated for all candidate regions of the genome. The candidate regions are defined as each region of the genome that is coding and has a length of 30 nucleotides. This was done using a sliding window approach. A window size of 30 nucleotides was chosen because, when translated, this corresponds to a peptide sequence of approximately 10 amino acids (depending on the ORF) and this is the size of the peptides that MHC-I can bind and present as antigens.

In order to investigate the phylogenetic relationship between the different variants the sequences were first clustered using CD-HIT (Li and Godzik, 2006). Afterwards FastTree was used to create phylogenetic trees of the representative sequences (Price et al., 2010).

A consensus sequence was made from the MSA in order to investigate the relationship between the secondary structure and conservation of the RNA. The secondary structure of the consensus was then predicted using the ViennaRNA package (Lorenz et al., 2011).

3. MATERIALS AND METHODS

3.1 Sequence retrieval

Influenza A virus

The IAV sequences were downloaded from the NCBI Influenza Virus Resource (Bao et al., 2008). All sequences of each of the eight genome segments were downloaded using the Unix command rsync and the file transfer protocol of NCBI. The IAV sequences on NCBI are not the genomic negative-sense sequences but their transcribed positive-sense counterparts (cRNA). Sequences that were not collected from human hosts or were incomplete were discarded.

```
$ rsync -aP --include=influenza\_na.dat.gz --include=influenza.fna.gz
--exclude=∗ rsync://ftp.ncbi.nlm.nih.gov/genomes/INFLUENZA/ [output file
    ]
```

—options uitleg?—-

SARS-CoV-2

All the full-length SARS-CoV-2 genome sequences from human hosts were downloaded from the NCBI Virus database using their Datasets command line tool (Hatcher et al., 2017).

```
$ datasets download virus genome taxon SARS-CoV-2 --host human --
    complete-only --exclude-cds --exclude-protein --filename [output file
    ]
```

—options uitleg?—-

3.2 Sequence preprocessing and filtering

For both viruses the same preprocessing was performed. Firstly, Sequences without collection date were filtered out. This was done because the collection date is important for calculating the mutability metric, which will be discussed in section 3.4.2. Secondly, sequences containing ambiguous characters (such as ’K’, ’V’ and ’H’ which signify uncertain nucleotides) were also removed since they skew the conservation calculations. Finally, duplicate sequences were removed as well to limit the size of the MSA which is the most time consuming step in this pipeline. If identical sequences were found, their collection dates were compared and the sequence identifier with the earliest collection date was kept. This is relevant for the calculation of the mutability. The IAV sequences were also separated and sorted according to which segment they belonged to. All subsequent steps of the pipeline and analyses were done separately for each of the segments. Note that the amount of sequences per segment differs.

3.3 Multiple sequence alignment

Influenza A virus

The sequences of the eight genome segments were aligned separately . This was done using the MAFFT command line tool (Katoh et al., 2002). MAFFT contains several different types of alignment algorithms, each optimized for different types of sequence datasets. The tool also has an ’–auto’ setting which automatically detects the optimal algorithm for the input dataset, this auto setting was used for the IAV alignment.

```
$ mafft --thread -1 --auto --preservecase [input fasta file] > [output
    MSA]
```

Setting ’–thread’ to -1 allows for faster, parallel computation if the computer has multiple processing cores. The option ’–preservecase’ is added to keep the sequences in capital letters.

SARS-CoV-2

The MAFFT auto setting was not appropriate for the SARS-CoV-2 dataset because its genome is significantly longer than the IAV genome segments and because there were also a lot more sequences (the amount of sequences used in the analysis are discussed in section 4.1). In order to deal with very large datasets of closely-related viral genomes MAFFT has specific setting that can be used.

```
$ mafft --thread -1 --auto --preservecase --6merpair --addfragments [
    input fasta file] [reference fasta file] > [output MSA]
```

In this mode, one sequence is supplied as reference and all others are aligned to this reference sequence. However this mode of action is still experimental (Research Institute for Microbial Diseases, Osaka University, 2021). The option ’–addfragments’ facilitates this behaviour. This is only accurate if the viral genomes are very closely-related, which is the case for the SARS-CoV-2. The option ’–6merpair’ allows for faster, albeit less accurate computation. The increased speed comes from estimating the distance matrix between two sequences using the amount of shared 6-mers. Even though these settings speed up the MSA significantly it still had to performed using the Central High Performance Computing Infrastructure of Ghent University (HPC) because of the large size of the dataset (Ghent University, nd).

3.3.1 Alignment trimming

Rare insertions form a problem when trying to score the conservation of a set of aligned sequences. They form columns in the alignment that are mostly made up of gaps (denoted with ’-’) and are thus sparse in actual nucleotide characters (’A’, ’T’, ’G’ and ’C’). Calculating the conservation of these these positions returns a high degree of conservation, since very little variation is observed. Biologically this is not true so therefor the rare insertions were trimmed out of the alignment.

The gap percentage was calculated for each position in the MSA. This is equal to the amount of sequences that contain a gap at this position divided by the total amount of sequences multiplied by 100%. Then, the average length of the unaligned sequences was calculated. The columns with a gap percentage lower then a certain threshold were kept in the alignment, the others trimmed away. The threshold was iteratively lowered (from 100% in steps of 1%) until the amount of columns in alignment was less than or equal to the average sequence length. This resulted in an alignment that had roughly the same length as the average sequence.

3.4 Conservation analysis

The conservation of each nucleotide position, corresponding to a column in the alignment, was scored using two metrics: the Shannon entropy and the mutability.

3.4.1 Shannon entropy

The Shannon entropy (H) is an often used metric in informatics that describes the uncertainty of the value of a variable, given the probabilities of all the values that the variable can take (Shannon, 1948). In the context of conservation scoring from a MSA, H can be roughly interpreted as the amount of variation at a certain position in the genome. The Shannon entropy for a certain position (x) in the alignment can be calculated using equation 3.1.

$$H(x) = -\sum_{c \in C} p_c \log_2 (p_c) \,. \tag{3.1}$$

In the case of a categorical variable, C is the vector of all possible values that the variable c can take. Here, C contains all the characters that can be present at a certain positions in the alignment, $C = [A, T, G, C, -]$. One character in C is denoted by c and the probability of that character occurring in position x is denoted by p_c. p_c is calculated as the frequency of character c in that column of the MSA. Note that when a certain character never occurs in a certain column the term $p_c \log_2 (p_c)$ for that character is set to 0 since the logarithm of 0 is not defined.

A high Shannon entropy value at a certain position indicates high variation and thus low conservation. The minimal value of H is 0 which occurs when $p_c = 1$ for one of the characters and $p_c = 0$ for all others. This corresponds to a column in the MSA being perfectly conserved and only containing one character. The maximal value of H is found when $p_c = 1/l$ for all $c \in C$ where l is the length of C. In this case the maximal value of H is approximately equal to 2.32 which corresponds to a column in the MSA where all five characters occur with equal frequency.

3.4.2 Mutability

—-info about why we wanted to use a second metric?—— —-explain both measures with a picture——

A second metric was devised to score the conservation at each position (x), the mutability (M). To calculate $M(x)$ the aligned sequences are first ordered according to their collection date. Then, starting from the earliest collected sequence, subsequent pairs of sequences are compared. Each time a character difference was found at a certain position this was recorded as a 'mutation' at that position (gaps were also included). Note that this does not signify an actual mutation event but is just a difference between two collected sequences. The sum of these mutations for one position over all subsequent sequence pairs ($m(x)$) was then divided by the total amount of sequences (n) in the MSA in order to normalize the measure. This is described by equation 3.2.

$$M(x) = \frac{m(x)}{n} \tag{3.2}$$

When a column in the alignment is completely conserved, only one character occurs and there are no mutations recorded so the mutability of that position is 0. The highest value of M is found when in each subsequent sequence a character mutation is observed. This corresponds to a value of $\frac{n-1}{n}$. The numerator is the amount of possible sequence pairs which is equal to the amount of sequences minus one. For a large amount of sequences this maximal mutability value is very close to 1.

3.4.3 Finding the most conserved regions

As mentioned before the goal is to look for the regions in the genome that have the highest total conservation. This was done using a sliding window over the sequence with a window size of 30 nucleotides. For the IAV this was performed segmentwise as to avoid windows spanning the gap between two segments. The total conservation of a window was calculated by taking the sum of the conservation metric over these 30 positions. This was done separately for the Shannon entropy and the mutability. For both metrics, low values indicate a high degree of conservation and vice versa.

Only coding sequences (CDS) of the genome can be translated into protein and presented as an antigen. Because of this the windows that are not fully contained inside a CDS have to be filtered out. To determine the location of the CDS in the trimmed MSA a consensus sequence was constructed. This was done by taking the most frequently occurring nucleotide in each position in the alignment (gaps were not included). The annotated CDS of each of the viruses were downloaded manually from the NCBI Nucleotide database website (Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information, 1988).

For the SARS-CoV-2 the CDS of the reference SARS-CoV-2 Wuhan-Hu-1 strain (Nucleotide accession NC_045512, version NC_045512.2) was used. For the IAV the Influenza A virus (A/California/07/2009(H1N1)) (RefSeq accession GCF_001343785.1) reference strain was used. The CDS had to be downloaded separately per segment, their accessions are displayed in table 3.1.

Then, each of the CDS were aligned to their respective consensus sequence using MAFFT with settings that align the shorter CDS to the longer consensus.

```
$ mafft --thread -1 --auto --preservecase --keeplength --addfragments [
    input CDS fasta file] [input consensus fasta file] > [output MSA]
```

Some of the CDS had insertions relative to the consensus sequence which creates gaps in the consensus. The ’–keeplength’ option was used to avoid this since this creates problems with the positioning of the 30 nucleotide regions. For the IAV the CDS of the M2 gene and the NEP gene had to be split into two parts since these are splicing variants and would not align properly otherwise. The 30 nucleotide regions were then filtered. The ones that were not fully included inside a CDS were removed and not considered as vaccine candidates. The remainder of the 30 nucleotide regions are hereinafter referred to as ’candidate regions’.

Table 3.1: The NCBI Nucleotide accessions and versions of the downloaded CDS for each of the Influenza A virus (A/California/07/2009(H1N1)) segments.

Segment	Accession	Version
1	NC_026438	NC_026438.1
2	NC_026435	NC_026435.1
3	NC_026437	NC_026437.1
4	NC_026433	NC_026433.1
5	NC_026436	NC_026436.1
6	NC_026434	NC_026434.1
7	NC_026431	NC_026431.1
8	NC_026431	NC_026431.1

3.5 Phylogenetic analysis

3.5.1 Sequence clustering

The sequences were clustered since the amount of sequences was too large to construct a phylogenetic tree. For this the cd-hit-est function of the CD-HIT software was used on the unaligned sequences (Li and Godzik, 2006). CD-HIT is specifically adapted for fast clustering of large datasets by using short word based heuristics and so avoiding full-length pairwise sequence alignment. For each cluster CD-HIT returns a representative sequence. The other sequences are at least as similar to this representative sequence as the identity threshold that is specified by the user.

```
$ cd-hit-est -i [input fasta file] -o [output cluster file] -c [identity
    threshold] -n 10 -T 0
```

The ’–n’ option controls the word length and depends on the identity threshold used. Setting ’–T’ to 0 allows for parallel computation. The sequences were clustered using several identity thresholds in order to determine the one that produced a reasonable amount of clusters for tree construction. The results of these clusterings are discussed in section 4.4.1. Clustering of the SARS-CoV-2 sequence had to be performed on the HPC.

3.5.2 Phylogenetic tree construction

The representative sequences returned by CD-HIT were used to build a phylogenetic tree. While clustering was performed on the unaligned sequences, tree construction was done using the untrimmed aligned sequences. To construct these trees the FastTree software was used (Price et al., 2010).

```
$ fasttree -nt -quote [input clustered MSA] > [output tree]
```

The ’–nt’ setting is used for nucleotide alignments. The ’–quote’ option allows for spaces in the fasta header. The trees were visualized using Dendroscope (Huson and Scornavacca, 2012).

3.6 Secondary structure analysis

3.6.1 Secondary structure prediction

The RNA secondary structure of the consensus was predicted from the consensus sequences using the RNAfold algorithm of the ViennaRNA package (Lorenz et al., 2011).

```
$ RNAfold -p0 -d2 --noPS --infile=[input consensus fasta file] --outfile
    =[output dotbracket structure file]
```

This algorithm returns the predicted structure of the RNA in dotbracket notation. The dotbracket notation uses dots (’.’) to represent unpaired nucleotides and brackets (’(’ and ’)’) to denote paired nucleotides in the RNA sequence. By using the ’–p0’ and ’–d2’ options the centroid structure is returned instead of the more commonly used used minimum free energy structure. Ding et al. (2005) have shown that the centroid structure is generally a better estimate of the structure derived from comparative sequence analysis than the minimum free energy structure. The centroid structure is based on all possible secondary structures of a certain RNA sequence, not just the one that has the least free energy. The ’–noPS’ setting was used to avoid making a figure of the secondary structure. The secondary structure prediction for the SARS-CoV-2 was performed on the HPC.

—-picture of sequence + dotbracket—-

Influenza A

The IAV secondary structures were visualised using the ViennaRNA Forna webapp (Kerpedjiev et al., 2015). Forna visualises the RNA secondary structure using the sequence and the dotbracket structure. It has several display options such as coloring the nucleotides. In order to create the figures in 4.5.1 the nucleotides were colored according to their mutability. In order to achieve a good color distribution the mutability was transformed using equation 3.3.

$$C(x) = \frac{\log_{10}(M(x) * 1000 + 1)}{\max(\log_{10}(M(x) * 1000 + 1))} \tag{3.3}$$

$C(x)$ is the color value for each position (x). $C(x)$ limited between 0 and 1 as is required by Forna.

SARS-CoV-2

The SARS-CoV-2 secondary structure was too large to display in the forna webapp so it was plotted using the RNAplot function of the ViennaRNA package. Here, the nucleotides could not be colored.

```
  $ RNAplot --infile=[input dotbracket structure file]
```

3.6.2 Link between secondary structure and conservation

Structure elements

To investigate the correlation between the secondary structure of the RNA and its conservation the structure elements had to be determined. This was done using the ViennaRNA Forgi library (Thiel et al., 2019). This library contains a script rnaConvert.py that can convert the dotbracket RNA structure into structure elements. The script can be used as a command line function. The ’–T’ option allows to specify to which format the structure should be converted.

```
$ rnaConvert.py [input dotbracket structure file] -T element_string --
    filename [output element structure file]
```

The Forgi package recognizes six types of structure elements:

- 5’ unpaired nucleotides (f)
- 3’ unpaired nucleotides (t)
- stem or helix (s)
- interior loop and bulge (i)
- multiloop (m)
- hairpin loop (h)

Each nucleotide in the sequence is then assigned to one of these elements by the algorithm based on the dotbracket structure. Regretfully the algorithm does not distinguish between interior loops and bulges and is also unable to recognize pseudoknots.

—-picture of sequence + dotbracket + element —-

Statistical analysis

Anova adjusted for differences in sample size, structure elements as explanatory variables for the conservation metric

4. RESULTS

4.1 Sequence retrieval, preprocessing and filtering

The conservation analysis required a clean sequence dataset. To this end the sequences were filtered. Firstly, in order to calculate the mutability the sequences were required to have a complete collection date. The sequences without one were discarded. Secondly, sequences with ambiguous characters were filtered out since they skew the calculation of both Shannon entropy and mutability. Finally, duplicate sequences were removed in order to reduce the size of the dataset. While these are quite stringent selection criteria, the resulting amount of sequences still remained relatively large which justifies this filtering a large extent.

Table 4.1: Overview of number of sequences before filtering and after each subsequent filtering step for the IAV and SARS-CoV-2.

Virus	Segment	Unfiltered	Date	Ambiguous	Duplicates
IAV	1	37111	34556	32545	20188
	2	36421	33826	21619	19603
	3	37378	34792	32793	19793
	4	53777	49567	46127	29435
	5	38357	35675	34406	16980
	6	47288	43561	40669	23232
	7	42567	39422	38338	13121
	8	38602	35806	34721	13133
SARS-CoV-2		1074845	1047425	466449	348678

Table 4.1 displays the reduction in the amount of sequences after each successive filtering step. As you can see the removal of sequences without collection date was relatively unimpactful. For the IAV the subsequent removal

of ambiguous sequences also did not reduce the amounts significantly except for segment 2 (reduction of 36% compared to the previous step). In contrast, the removal of ambiguous SARS-CoV-2 sequences lead to a large reduction of 55%. Finally, the filtering of duplicates did lower the amount of sequences quite significantly with an average reduction of 45% for the IAV and 25% for the SARS-CoV-2 compared to the previous step.

4.2 Alignment trimming

As can be seen in table 4.2 the untrimmed MSA was significantly longer then the average sequence. This was mostly because of rare insertions, corresponding to highly gapped columns in the alignment. These columns interfere with the conservation analysis due artificially high conservation values which are not relevant. In order to remedy this the MSA was trimmed until it had a length similar to the average unaligned sequence. This was done using a gap percentage threshold which is also displayed in table 4.2. Alignment columns with a gap percentage higher than this threshold were trimmed away.

Table 4.2: Overview of the average sequence length, the length of the untrimmed and trimmed MSA and the gap percentage threshold used during trimming for the IAV and SARS-CoV-2.

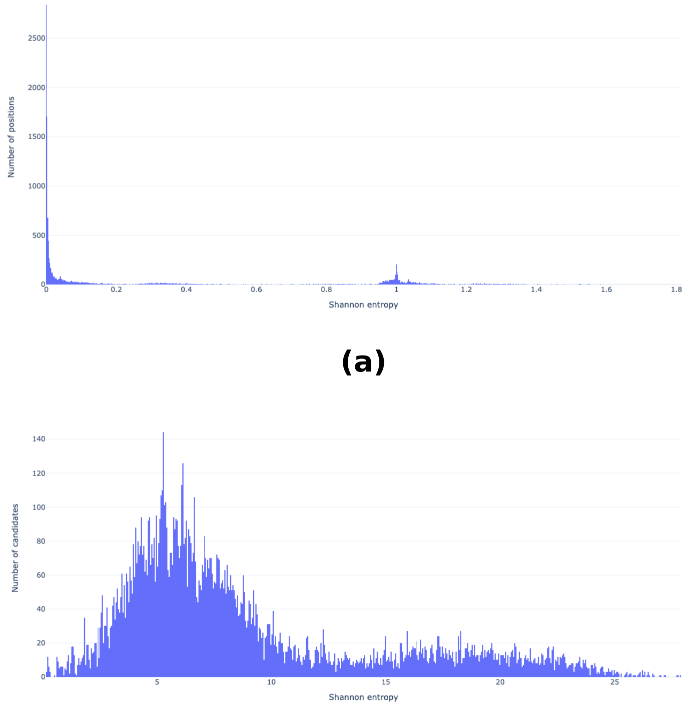
Virus	Segment	Average	Untrimmed	Trimmed	Threshold(%)
IAV	**1**	2303	2692	2297	38
	2	2303	2447	2303	34
	3	2189	2306	2189	39
	4	1724	2283	1718	52
	5	1527	1817	1526	38
	6	1426	1877	1426	51
	7	994	1266	994	48
	8	857	990	856	40
SARS-CoV-2		29784	32164	29783	45

4.3 Conservation analysis

During the conservation analysis the Shannon entropy and mutability were calculated for each position in the MSA. These metrics were then calculated for each 30 nucleotide window by summing them over this window. The windows that were not entirely contained in a CDS were filtered out, leaving us with all potential vaccine candidate regions. The candidate regions that obtained the lowest scores for both metrics are the ones that are the most conserved and thus have the most potential as a universal vaccine.

4.3.1 Influenza A

Shannon entropy



(a)

(b)

Figure 4.1: Histograms of the Shannon entropy for the IAV. The data from all genomic segments was combined to make these figures. (a) The histogram for the genomic positions. (b) The histogram for the candidate regions.

Figure 4.1 clearly shows the bimodal distribution of the Shannon entropy for both the positions in the genome and for the candidate regions. In figure 4.1a there is a clear overrepresentation of Shannon entropy values close to zero. This means that most of the positions in the genome are relatively conserved. The second peak is smaller and located near the Shannon entropy value of one. This can be explained by columns in the MSA that are dominated by two characters in approximately equal proportions. These are positions where mostly only two nucleotides (or a gap and a nucleotide) occur with roughly the same frequency.

Figure 4.1b also displays a bimodal distribution. This comes from the fact that segment four and seven had significantly higher average Shannon entropy values which is illustrated in figure 4.2. This is not surprising because these segments code for the HA and NA proteins which are known to be more variable (Plotkin and Dushoff, 2003). Compared to 4.1a the Shannon entropy values are higher since these values represent the sum over a window of 30 nucleotides. You can see that the largest peak is not located around a Shannon entropy of zero anymore. The amount of candidate regions with a cumulative Shannon entropy close to zero is actually very limited. This means that while a there are many conserved positions in the genome, there are a lot less conserved stretches of positions.

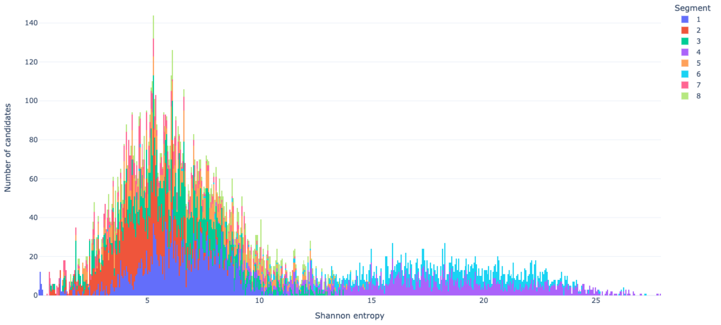


Figure 4.2: Histogram of the Shannon entropy for each candidate region of the IAV. Each of the genomic segment is colored differently.

—-table of average Shannon entropy per segment/protein CDS?—-

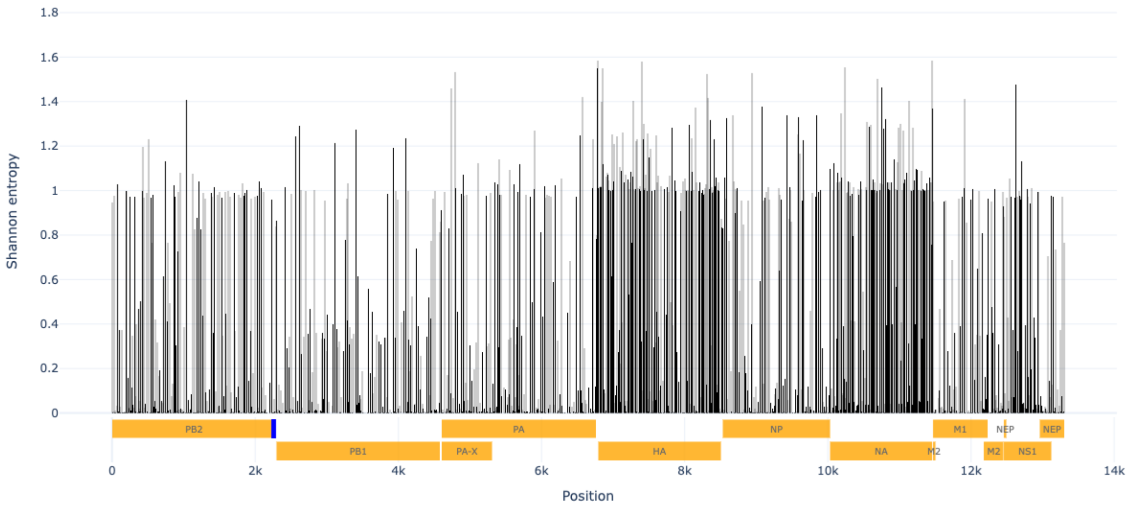
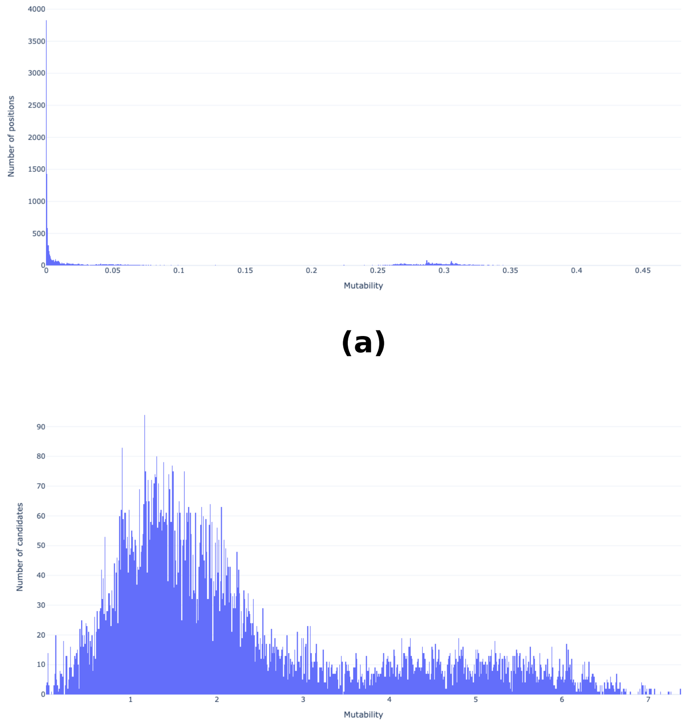


Figure 4.3: Overview of the Shannon entropy, CDS and most conserved regions of the entire IAV genome. The black vertical bars display the Shannon entropy for each position in the genome. The yellow horizontal bars are the CDS after alignment to the consensus sequence. The small blue bars indicate the ten most conserved candidate regions according to the Shannon entropy. Here, these all cluster together in one region.

Figure 4.3 displays Shannon entropy of the entire IAV genome and its CDS. You tell by the darker regions and higher bars in the Shannon entropy graph that segment four and six are more variable then the others. The blue bars indicate the ten most conserved candidate regions. These all cluster together near the 3' end of the PB2 CDS which means that this is the most conserved part of the IAV genome according to the Shannon entropy. An overview of these top ten most conserved sequences is given in table 4.3.

Mutability

The results of the conservation scoring using the mutability metric are extremely similar to the the ones using the Shannon metric and will thus not be discussed in detail. The mutability histograms in figure 4.4 show a similar bimodal distribution as for the Shannon entropy. Figure 4.5 shows that the top ten most conserved regions are again located in the 3' region of PB2. This top ten is actually exactly the same as for the Shannon entropy but in a slightly different order, I refer again to table 4.3 for these sequences.



(a)

(b)

Figure 4.4: Histograms of the mutability for the IAV. The data from all genomic segments was combined to make these figures. (a) The histogram for the genomic positions. (b) The histogram for the candidate regions.

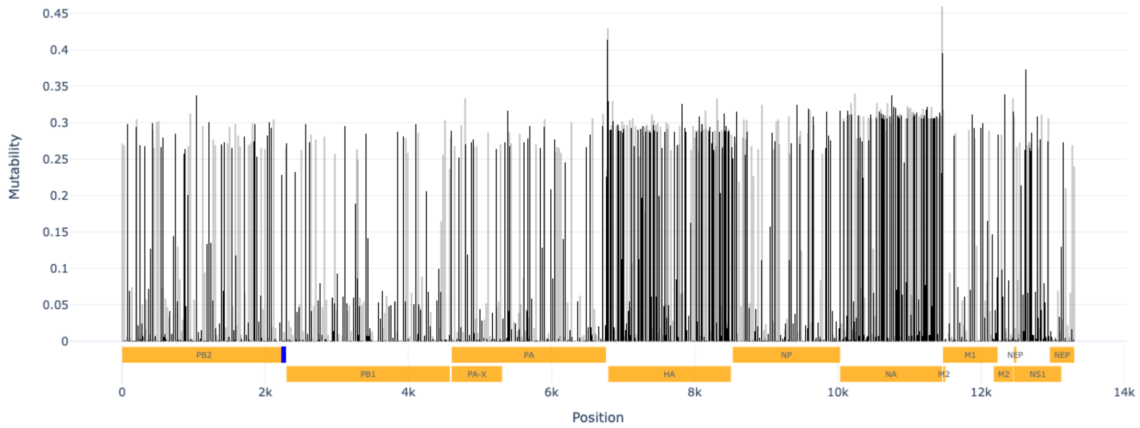


Figure 4.5: Overview of the mutability, CDS and most conserved regions of the IAV genome. The black vertical bars display the mutability for each position in the genome. The yellow horizontal bars indicate the CDS after alignment to the consensus sequence. The small blue bars are the ten most conserved candidate regions according to the mutability. Here, these all cluster together in one region.

Table 4.3: Overview of the ten most conserved regions in the IAV genome according to the Shannon entropy and mutability metrics. The sequences are ranked from most to least conserved according to the Shannon entropy. The positions are defined as compared to the consensus sequence that was constructed from the trimmed MSA. These are the positions on the entire genome as well as the positions on the segment, since these sequences all lie in the first segment. The Shannon entropy and mutability displayed here are the summed values over all 30 positions of the sequence.

Ranking	Positions	Sequence	Shannon entropy	Mutability
1	2234-2263	CTTACTGACAGCCAGACAGCGACCAAAAGA	0.15754	0.02912
2	2232-2261	TACTTACTGACAGCCAGACAGCGACCAAAA	0.15832	0.02923
3	2233-2262	ACTTACTGACAGCCAGACAGCGACCAAAAG	0.15910	0.02932
4	2247-2276	AGACAGCGACCAAAAGAATTCGGATGGCCA	0.20272	0.03625
5	2248-2277	GACAGCGACCAAAAGAATTCGGATGGCCAT	0.20428	0.03646
6	2249-2278	ACAGCGACCAAAAGAATTCGGATGGCCATC	0.20888	0.03725
7	2235-2264	TTACTGACAGCCAGACAGCGACCAAAAGAA	0.21972	0.04072
8	2250-2279	CAGCGACCAAAAGAATTCGGATGGCCATCA	0.22430	0.03963
9	2251-2280	AGCGACCAAAAGAATTCGGATGGCCATCAA	0.22780	0.04012
10	2252-2281	GCGACCAAAAGAATTCGGATGGCCATCAAT	0.22936	0.04032

Correlation between Shannon entropy and mutability

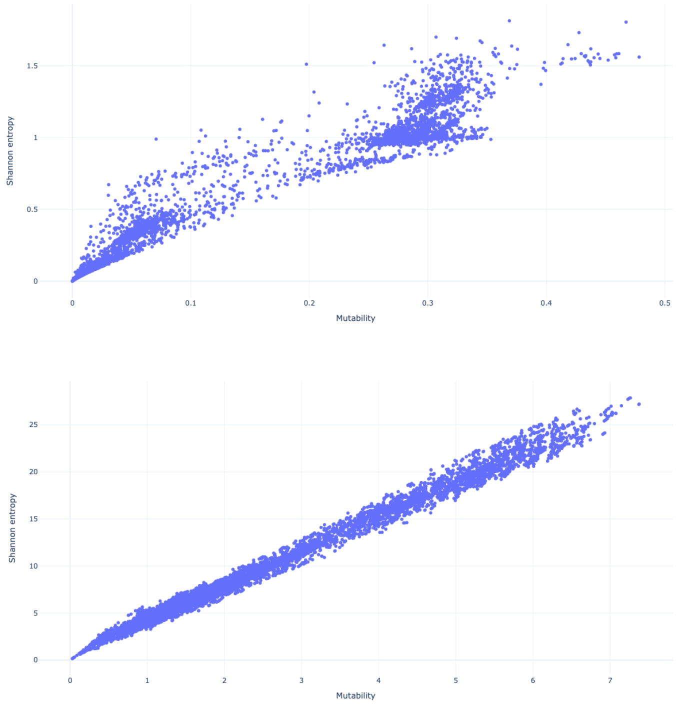


Figure 4.6: Scatter plots of the Shannon entropy and the mutability for the IAV. The top graph shows this for each position in the genome. The bottom graph shows this for every candidate region.

It is clear from the previous results and from figure 4.6 that there seems to be a strong correlation between the Shannon entropy and mutability for the IAV. The correlation seems to be stronger for the candidate regions then for the single positions. This is likely be due to an averaging effect because of the summing of the metrics over a window.

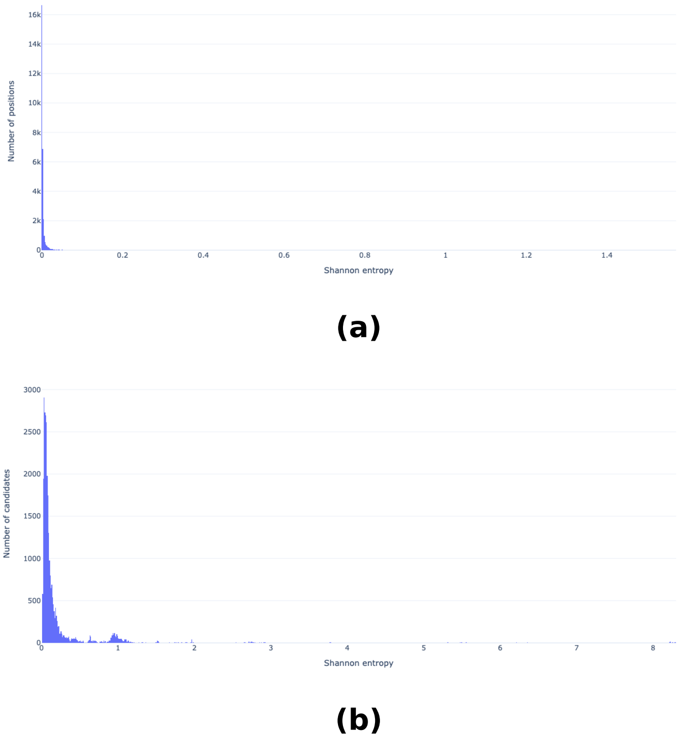
—-R^2 values?—-

4.3.2 SARS-CoV-2

Shannon entropy

The distribution of the Shannon entropy for the SARS-CoV-2 is quite different from the IAV as can be seen in figure 4.7. For both the positions and candidate regions the large majority is located around low Shannon entropy values with a clear peak near zero. There is also no significant second peak.

This indicates that there is less variation in the SARS-CoV-2 dataset compared to the IAV dataset. This could lead to think that the SARS-CoV-2 genome might be more conserved than the IAV genome. This idea is supported by the fact that the SARS-CoV-2 has a form of proofreading during replication due to its nsp14 exonuclease domain (Sola et al., 2015). The IAV is also generally reported to have a higher mutation rate but this has also been disputed () *****REF*****. However, we should be very careful with generalizing this conclusion since there is a large temporal and size imbalance between the SARS-CoV-2 and IAV virus datasets, this will be discussed further in section 5.



(a)

(b)

Figure 4.7: Histograms of the Shannon entropy for the SARS-CoV-2. (a) The histogram for the genomic positions. (b) The histogram for the candidate regions.

Figure 4.8 displays the entire SARS-CoV-2 genome with its CDS and the Shannon entropy per position. Again it is noticeable that in general the Shannon entropy is lower than for the IAV. When calculated, the average Shannon entropy over all positions of the IAV genome was equal to 0.3 compared to 0.01 for the SARS-CoV-2. For SARS-CoV-2 the ten most conserved candidate regions don't all cluster in the same region of the genome. The most conserved region lies in the S protein CDS. This is relatively surprising since the S protein is known to have a higher mutation rate then the over-all genome (Amicone et al., 2022; Wang et al., 2022). Two of the top ten regions lie in the E protein CDS and the other seven in the ORF1ab.

——-ORF1ab proteins toevoegen op figuur——

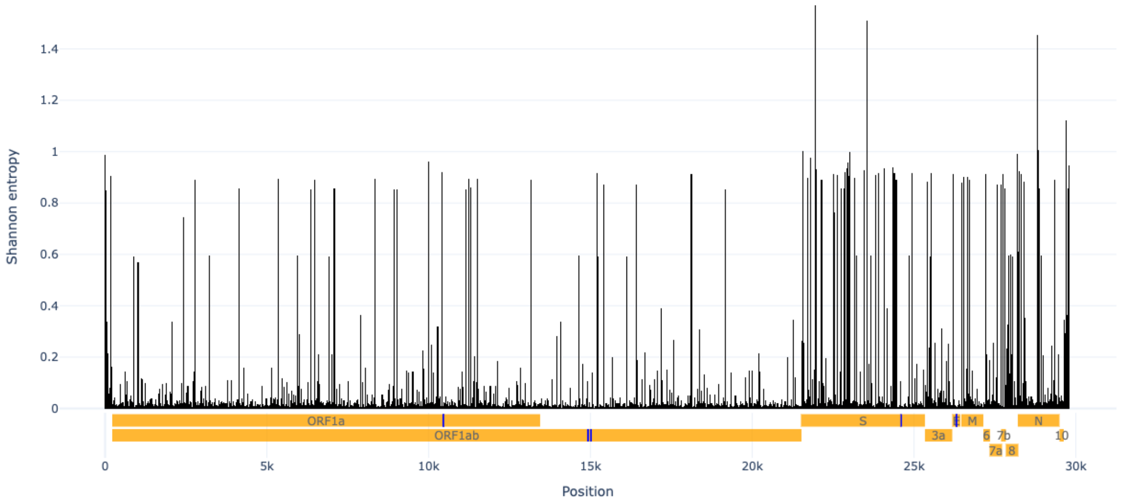
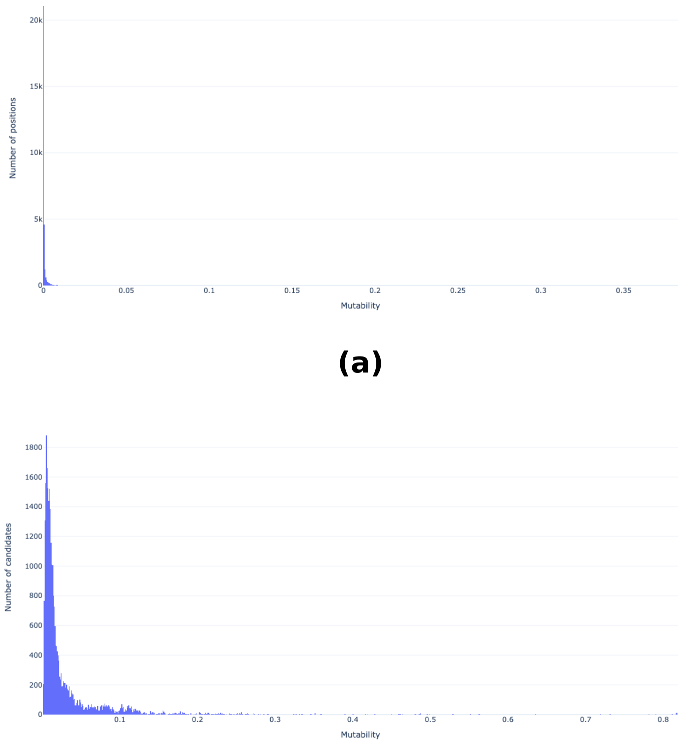


Figure 4.8: Overview of the Shannon entropy, CDS and most conserved regions of the SARS-CoV-2 genome. The black vertical bars display the Shannon entropy for each position in the genome. The yellow horizontal bars indicate the CDS of the genome after alignment to the consensus sequence. The small blue bars indicate the ten most conserved candidate regions according to the Shannon entropy.

Mutability

The mutability distributions are again quite similar compared to the Shannon entropy distributions with one major peak near zero, as can be seen in figure 4.9. This again indicates a strong overrepresentation of conserved positions and even 30 nucleotide stretches of positions in the SARS-CoV-2 genome.

Figure 4.10 displays the ten most conserved candidate regions according to the mutability. There is large overlap compared to the Shannon entropy top ten but there are some differences.



(a)

(b)

Figure 4.9: Histograms of the mutability for the SARS-CoV-2. (a) The histogram for the genomic positions. (b) The histogram for the candidate regions.

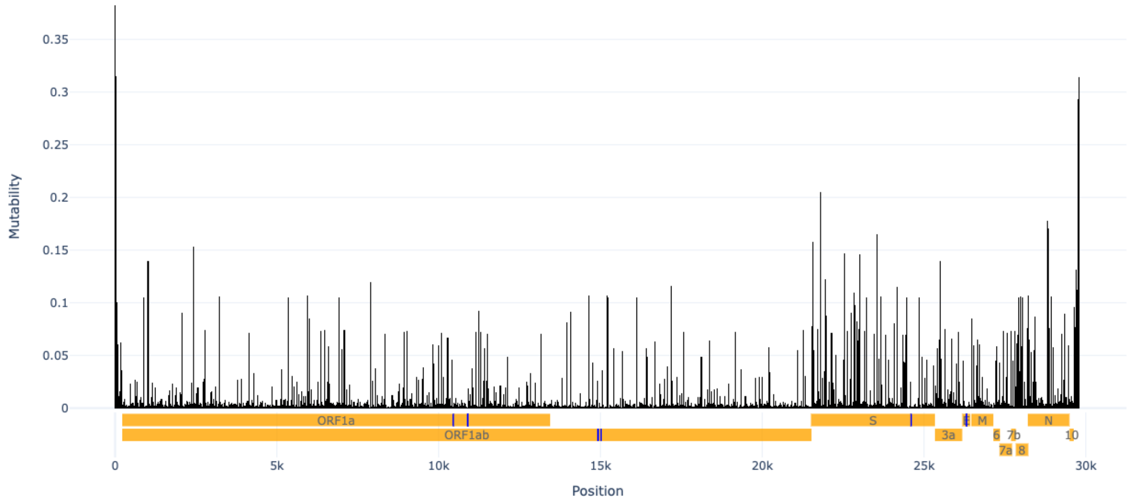


Figure 4.10: Overview of the mutability, CDS and most conserved regions of the SARS-CoV-2 genome. The black vertical bars display the mutability for each position in the genome. The yellow horizontal bars indicate the CDS of the genome after alignment to the consensus sequence. The small blue bars indicate the ten most conserved candidate regions according to the mutability.

Correlation between Shannon entropy and mutability

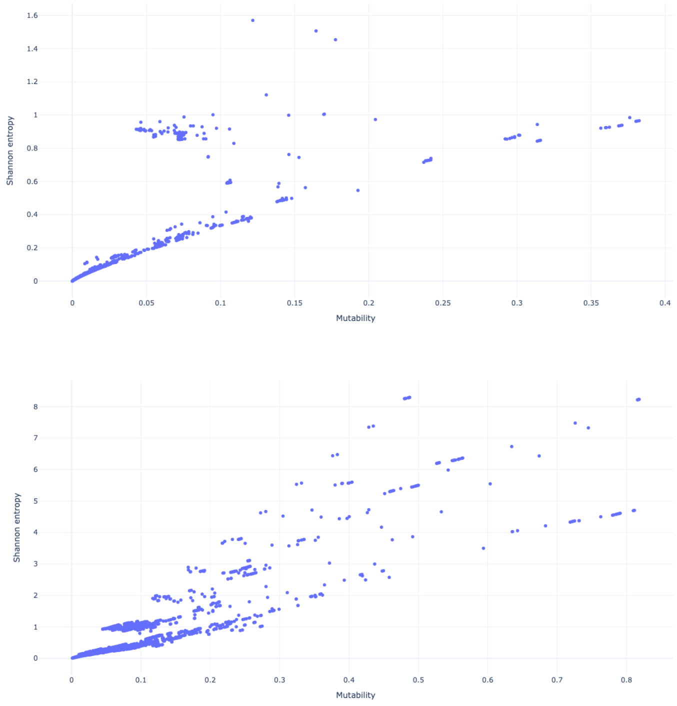


Figure 4.11: Scatter plots of the Shannon entropy and the mutability for the SARS-CoV-2. The top graph shows this for each position in the genome. The bottom graph shows this for every candidate region.

4.4 Phylogenetic analysis

4.4.1 Clustering

Influenza A

Table 4.4

Segment	**Identity threshold (%)**					
	99.5	**99**	**98**	**97**	**96**	**95**
1	1078	267	100	60	40	31
2	1085	231	86	48	30	23
3	1062	243	93	46	38	27
4	2750	503	141	87	63	55
5	1083	238	70	44	33	21
6	2187	565	143	89	56	47
7	1078	189	61	37	23	17
8	1216	313	84	54	37	29

SARS-CoV-2

Table 4.5

Identity threshold (%)						
99.95	**99.90**	**99.85**	**99.80**	**99.75**	**99.50**	**99**
14684	1117	301	185	143	81	57

4.4.2 Phylogenetic tree

Grouping of same subspecies & years

Influenza A

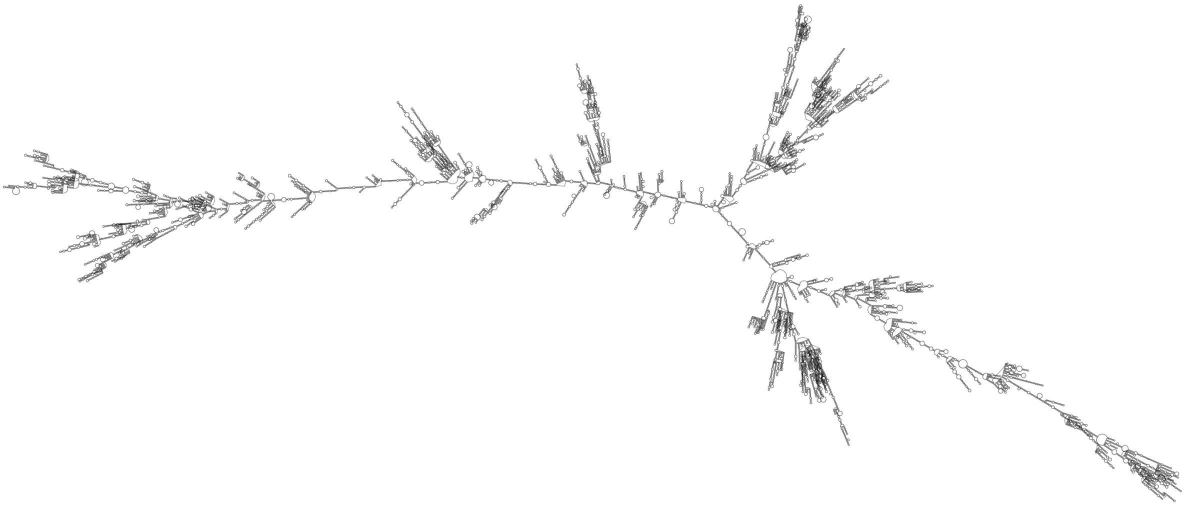
SARS-CoV-2

4.5 Secondary structure analysis

4.5.1 Secondary structure prediction

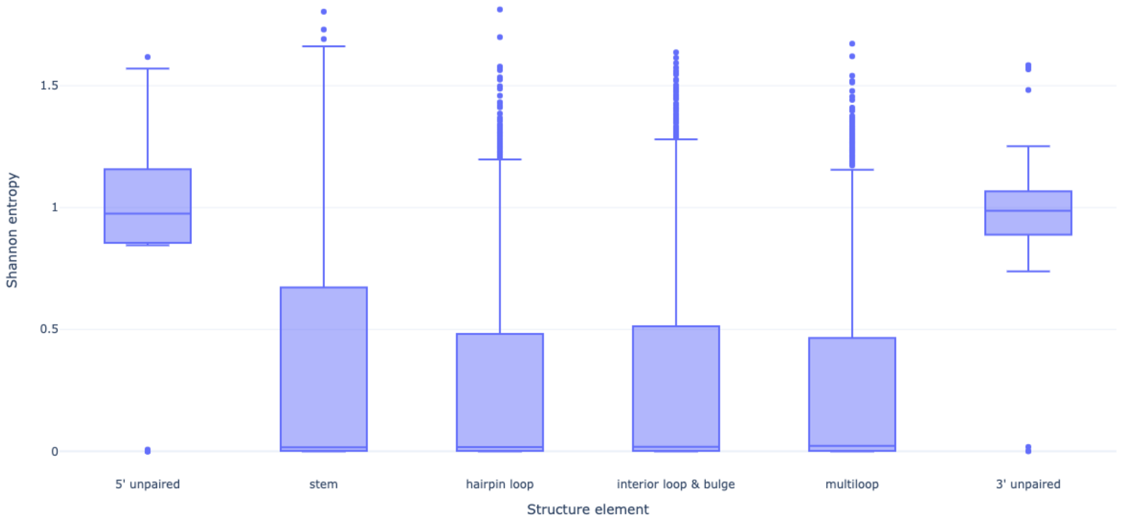
Influenza A

SARS-CoV-2

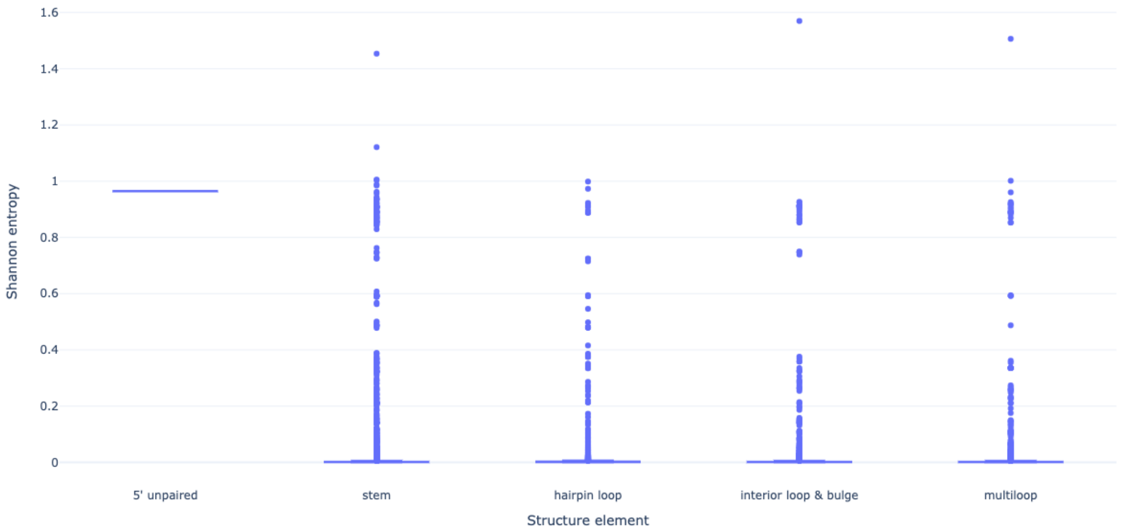


4.5.2 Link between secondary structure and conservation

Influenza A



SARS-CoV-2



5. DISCUSSION

Reflect on goals & objectives

Limitations of the study

- harsh filtering, no imputation of ambiguous nucleotides

- silent mutations

- dataset imbalance for sequence counts and variants - temporal dataset imbalance: way more SARS-CoV-2 sequences but collected over way less time

- assumptions of mutability (linear evolution)

- no immunogenicity prediction

What can be done in the future?

- other hosts, related viruses

- immunogencitiy prediction on found regions

- co-mutation analysis

- structure variation analysis

BIBLIOGRAPHY

Amicone, M., Borges, V., Alves, M. J., Isidro, J., Zé-Zé, L., Duarte, S., Vieira, L., Guiomar, R., Gomes, J. P., and Gordo, I. (2022). Mutation rate of sars-cov-2 and emergence of mutators during experimental evolution. *Evolution, medicine, and public health*, 10(1):142–155.

Bai, C., Zhong, Q., and Gao, G. F. (2021). Overview of sars-cov-2 genome-encoded proteins. *Science China Life Sciences*, pages 1–15.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., and Lipman, D. (2008). The influenza virus resource at the national center for biotechnology information. *Journal of virology*, 82(2):596–601.

Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (1988). Nucleotide [internet]. Retrieved May 26, 2022 from from: https://www.ncbi.nlm.nih.gov/nucleotide/.

Brian, D. and Baric, R. (2005). Coronavirus genome structure and replication. *Coronavirus replication and reverse genetics*, pages 1–30.

Byrd-Leotis, L., Cummings, R. D., and Steinhauer, D. A. (2017). The interplay between the host receptor and influenza virus hemagglutinin and neuraminidase. *International journal of molecular sciences*, 18(7):1541.

Calder, L. J., Wasilewski, S., Berriman, J. A., and Rosenthal, P. B. (2010). Structural organization of a filamentous influenza a virus. *Proceedings of the National Academy of Sciences*, 107(23):10685–10690.

Cheung, T. K. and Poon, L. L. (2007). Biology of influenza a virus. *Annals of the New York Academy of Sciences*, 1102(1):1–25.

De Vlugt, C., Sikora, D., and Pelchat, M. (2018). Insight into influenza: a virus cap-snatching. *Viruses*, 10(11):641.

Ding, Y., Chan, C. Y., and Lawrence, C. E. (2005). Rna secondary structure prediction by centroids in a boltzmann weighted ensemble. *Rna*, 11(8):1157–1166.

Dou, D., Revol, R., Östbye, H., Wang, H., and Daniels, R. (2018). Influenza a virus cell entry, replication, virion assembly and movement. *Frontiers in immunology*, 9:1581.

Eisfeld, A. J., Neumann, G., and Kawaoka, Y. (2015). At the centre: influenza a virus ribonucleoproteins. *Nature Reviews Microbiology*, 13(1):28–41.

Gamblin, S. J. and Skehel, J. J. (2010). Influenza hemagglutinin and neuraminidase membrane glycoproteins. *Journal of biological chemistry*, 285(37):28403–28409.

Ghedin, E., Sengamalay, N. A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D. J., Sitz, J., Koo, H., Bolotov, P., et al. (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, 437(7062):1162–1166.

Ghent University (n.d.). High performance computing infrastructure. Retrieved May 21, 2022 from https://www.ugent.be/hpc/en.

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., et al. (2020). A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature*, 583(7816):459–468.

Harris, H. and Hill, C. (2021). A place for viruses on the tree of life. *Frontiers in Microbiology*, 11:3449.

Hatcher, E. L., Zhdanov, S. A., Bao, Y., Blinkova, O., Nawrocki, E. P., Ostapchuck, Y., Schäffer, A. A., and Brister, J. R. (2017). Virus variation resource–improved response to emergent viral outbreaks. *Nucleic acids research*, 45(D1):D482–D490.

Huson, D. H. and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology*, 61(6):1061–1067.

Jin, H., Leser, G. P., Zhang, J., and Lamb, R. A. (1997). Influenza virus hemagglutinin and neuraminidase cytoplasmic tails control particle shape. *The EMBO journal*, 16(6):1236–1247.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066.

Kerpedjiev, P., Hammer, S., and Hofacker, I. L. (2015). Forna (force-directed rna): simple and effective online rna secondary structure diagrams. *Bioinformatics*, 31(20):3377–3379.

Khaperskyy, D. A., Schmaling, S., Larkins-Ford, J., McCormick, C., and Gaglia, M. M. (2016). Selective degradation of host rna polymerase ii transcripts by influenza a virus pa-x host shutoff protein. *PLoS pathogens*, 12(2):e1005427.

Krug, R. M. (2015). Functions of the influenza a virus ns1 protein in antiviral defense. *Current opinion in virology*, 12:1–6.

Lamb, R. A., Zebedee, S. L., and Richardson, C. D. (1985). Influenza virus m2 protein is an integral membrane protein expressed on the infected-cell surface. *Cell*, 40(3):627–633.

Leser, G. P. and Lamb, R. A. (2005). Influenza virus assembly and budding in raft-derived microdomains: a quantitative analysis of the surface distribution of ha, na and m2 proteins. *Virology*, 342(2):215–227.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). Viennarna package 2.0. *Algorithms for molecular biology*, 6(1):1–14.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The lancet*, 395(10224):565–574.

Mamuye, A., Merelli, E., and Tesei, L. (2016). A graph grammar for modelling rna folding. *Electronic Proceedings in Theoretical Computer Science*, 231:31–41.

Manrubia, S. C. and Lázaro, E. (2006). Viral evolution. *Physics of Life Reviews*, 3(2):65–92.

Melén, K., Fagerlund, R., Franke, J., Köhler, M., Kinnunen, L., and Julkunen, I. (2003). Importin α nuclear localization signal binding sites for stat1, stat2, and influenza a virus nucleoprotein. *Journal of Biological Chemistry*, 278(30):28193–28200.

Mitnaul, L. J., Castrucci, M. R., Murti, K. G., and Kawaoka, Y. (1996). The cytoplasmic tail of influenza a virus neuraminidase (na) affects na incorporation into virions, virion morphology, and virulence in mice but is not essential for virus replication. *Journal of virology*, 70(2):873–879.

Neumann, G., Hughes, M. T., and Kawaoka, Y. (2000). Influenza a virus ns2 protein mediates vrnp nuclear export through nes-independent interaction with hcrm1. *The EMBO journal*, 19(24):6751–6758.

of the International Committee on Taxonomy of Viruses, C. S. G. (2020). The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2. *Nature microbiology*, 5(4):536.

Paterson, D. and Fodor, E. (2012). Emerging roles for the influenza a virus nuclear export protein (nep). *PLoS pathogens*, 8(12):e1003019.

Payne, S. (2017). *Viruses: from understanding to investigation*. Academic Press.

Plotkin, J. B. and Dushoff, J. (2003). Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza a virus. *Proceedings of the National Academy of Sciences*, 100(12):7152–7157.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490.

Rampersad, S. and Tennant, P. (2018). Replication and expression strategies of viruses. *Viruses*, page 55.

Research Institute for Microbial Diseases, Osaka University (January 29 2021). Rapid calculation of full-length msa of closely-related viral genomes. Retrieved May 20, 2022 from https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html.

Rossman, J. S., Jing, X., Leser, G. P., and Lamb, R. A. (2010). Influenza virus m2 protein mediates escrt-independent membrane scission. *Cell*, 142(6):902–913.

Rossman, J. S. and Lamb, R. A. (2011). Influenza virus assembly and budding. *Virology*, 411(2):229–236.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Federhen, S., et al. (2010). Database resources of the national center for biotechnology information. *Nucleic acids research*, 39(suppl_1):D38–D51.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Sola, I., Almazan, F., Zúñiga, S., and Enjuanes, L. (2015). Continuous and discontinuous rna synthesis in coronaviruses. *Annual review of virology*, 2:265–288.

Suarez, D. L. (2016). Influenza a virus. *Animal Influenza*, pages 1–30.

Synowiec, A., Szczepański, A., Barreto-Duran, E., Lie, L. K., and Pyrc, K. (2021). Severe acute respiratory syndrome coronavirus 2 (sars-cov-2): a systemic infection. *Clinical microbiology reviews*, 34(2):e00133–20.

Thiel, B. C., Beckmann, I. K., Kerpedjiev, P., and Hofacker, I. L. (2019). 3d based on 2d: Calculating helix angles and stacking patterns using forgi 2.0, an rna python library centered on secondary structure elements. *F1000Research*, 8.

Varga, Z. T. and Palese, P. (2011). The influenza a virus protein pb1-f2: killing two birds with one stone? *Virulence*, 2(6):542–546.

Wang, M.-Y., Zhao, R., Gao, L.-J., Gao, X.-F., Wang, D.-P., and Cao, J.-M. (2020). Sars-cov-2: structure, biology, and structure-based therapeutics development. *Frontiers in cellular and infection microbiology*, page 724.

Wang, S., Xu, X., Wei, C., Li, S., Zhao, J., Zheng, Y., Liu, X., Zeng, X., Yuan, W., and Peng, S. (2022). Molecular evolutionary characteristics of sars-cov-2 emerging in the united states. *Journal of medical virology*, 94(1):310–317.

Yadav, R., Chaudhary, J. K., Jain, N., Chaudhary, P. K., Khanra, S., Dhamija, P., Sharma, A., Kumar, A., and Handu, S. (2021). Role of structural and non-structural proteins and therapeutic targets of sars-cov-2 for covid-19. *Cells*, 10(4):821.