Identifying functional k-mers in genome sequences

Vincent Wilantara

A thesis presented for the degree of
Bachelor of Science

Algorithmic Bioinformatics
Heinrich Heine University Düsseldorf
Germany
27th June, 2022

Acknowledgments

I am extremely grateful to Prof. Dr. Gunnar Klau for giving me the opportunity to write this thesis as well as his generous support and extensive feedback. Many thanks to Prof. Alex Dilthey for being the second assessor and special thanks to Philipp Spohr for assisting me with Snakemake pipelines and providing great insights to my many questions.

Abstract

The quest to uncover the meaning of genomes has shone the spotlight onto k-mers. As a result, there has been extensive research done on decoding the linkage between k-mers and their impacts on biological functions. An interesting method that we found for determining this linkage was to conduct linguistic analyses on the genome, which draw analogies between keywords found in conventional texts and functional k-mers in genome sequences. By measuring the clustering level of k-mers and their enrichment levels in important genomic regions, a correlation can be made between k-mer clustering and functional elements, thus implying that the function of the genomic region is tied to those k-mer clusters. In this thesis, we replicated this method of identifying functional k-mers in genome sequences and applied it to human and coronavirus genomes.

Contents

1 Introduction

The genome is the cumulative sum of an organism's hereditary materials. Goldman and Landweber 2016 describes them as "the information repository of an organism". A genome sequence may be represented as a one-dimensional sequence of nucleotides (A, C, G, and T), and a segment of k consecutive nucleotides is called a k-mer. K-mers are widely used in computational genomics and sequence analysis, performing a critical role in the elucidation of genome functions. For instance, Chae et al. 2013 reported that certain k-mers characterize CpG island sequences in humans which are known to regulate gene expression through transcriptional silencing of genes (Deaton and Bird 2011, Jacinto and Esteller 2007). Therefore, the discovery and extrapolation of functional k-mers in genomes have been a major topic of research in the scientific community. In this thesis, we proposed and applied a method for finding important (functional) k-mers in genome sequences using linguistic analyses.

In Ortuño et al. 2007, it was discovered that keywords in literary texts show a strong clustering along the text, whereas common words are randomly distributed. Keywords are used for topic detection and provide contextual and semantic meaning. Similarly, Durand and Sankoff 2002 found that k-mers that have well-defined biological functions, as well as genome elements such as genes and exons, are spatially clustered along the genome. This potentially indicated that genome sequences are structured similarly to conventional texts and therefore behave in the same way. This hypothesis is supported by Hackenberg, Carpena, et al. 2011, who stated that such spatial clustering often translates into genome structures with a clear functional and/or evolutionary meaning. To better portray this analogy, a k-mer can thereby be referred to as a DNA "word" of length k, and their clustering level can be regarded as a measurement of importance in biological sequences.

This thesis will look at the method and procedure first introduced by Hackenberg, Rueda, et al. 2012, in which they developed an algorithm to compute the clustering coefficients of k-mers in the genomes of humans and mice using the fluctuations of distances between consecutive k-mers. They observed that DNA words (k-mers) with strong clustering in the genome elements are also highly enriched inside them, this implies that strongly clustered k-mers are more likely to be associated with the biological function of the genome element. This led to the discovery of a novel way of identifying functional k-mers in DNA, and the replication and implementation of this procedure is the primary focus of this thesis. The procedure will be applied on the chromosome 1 sequence of the latest human genome assembly hg38, primarily on exons. In light of the ongoing COVID-19 pandemic, we contemplated the viability of such an approach on viral genomes. A potential dictionary of diagnostic k-mers can be built using this method, and the successful identification of functional k-mers can help improve research on viral vaccines. Hence, we conducted experiments on the SARS-CoV and SARS-CoV-2 genomes to investigate this hypothesis.

In Section 2, we will give a brief introduction to the context and relevancy of our experiments with the SARS-CoV and SARS-CoV-2 genomes. The complete procedure, which includes enrichment analyses and the word clustering algorithm, which we will refer to as the HCR algorithm (Hackenberg, Carpena, Rueda), as well as the resources used in this thesis will be discussed in-depth in Section 3. In Section 4 we will be showcasing the results on the human genome hg38, as well as SARS-CoV and SARS-CoV-2 genomes.

Lastly, it is important to note that the fundamental concepts that will be discussed in this thesis such as word clustering and word enrichment should be understood as k-mer clustering and k-mer enrichment. This naming convention stems from the linguistic models that the procedure was derived from. This includes the terminologies "DNA words" and "words", they should, unless specified, mean "k-mers". Throughout this thesis, "k-mers" and "words" will be used interchangeably.

2 Background Information

2.1 SARS-CoV and SARS-CoV-2

In December 2019, cases of patients suffering from a new type of respiratory disease started to emerge in Wuhan, Hubei the People's Republic of China. Symptoms of the disease included fever, sore throat, and respiratory distress which was diagnosed to be caused by a novel coronavirus (nCoV-2019). It was later discovered that the virus shared a genetic resemblance to an earlier known coronavirus SARS-CoV and was renamed SARS-CoV-2. The disease was highly infectious and rapidly spreads across the globe. In March 2020, the World Health Organization declared the coronavirus disease 2019 (COVID-19) a pandemic. As of June 2022, the COVID-19 pandemic mounts over 500 million infected cases including 6 million deaths worldwide. Therefore, accelerating research and innovations to combat SARS-CoV-2 is a contemporary task, and the identification of important k-mers can be used for this purpose. For instance, Ali et al. 2021 used a k-mer based sequence representation to classify SARS-CoV-2 variants. They found that mutations of certain amino acids and amino acid positions in the spike S protein are responsible for some variations of the SARS-CoV-2. Our quest of identifying functional k-mers in genome sequences plays a similar role to this, as it allows the possibility of discovering diagnostic k-mers which can be used to distinguish genome elements in the viral genome. SARS-CoV will be used in the experiments primarily as a comparison.

2.2 Genome Structure

Due to time constraints, we needed to selectively decide on which genome element we deemed to have more functional or evolutionary significance. To this end, we need to better understand the genome structure of the coronaviruses. Coronaviruses are spherical, enveloped, single-stranded RNA viruses with multiple spike surface glycoproteins projected on their surface, forming their defining corona appearance. Among the RNA viruses, coronaviruses have the largest genome size. The genome size of SARS-CoV-2 is approximately 30.000 base in length and it shares around 88% sequence identity with that of SARS-CoV (Wu et al. 2020). There have been rigorous studies on understanding key stages of the SARS-CoV-2 life cycles as well as identifying key protein structures. Figure 1 illustrates the genome structure for both SARS-CoV and SARS-CoV-2.

Two classifications of proteins can be found in both coronavirus genomes, non-structural proteins (nsps) and structural proteins. There are 16 nsps and 4 major structural proteins present in both SARS-CoV and SARS-CoV-2 that are considered to be the main contributors that run the RNA machinery. The overall mechanism of these nsps and major structural proteins has been decoded and well understood by researchers such as Romano et al. 2020 and Arya et al. 2020, and is outlined as follows. From the 5' end, two main ORFs (Open Reading

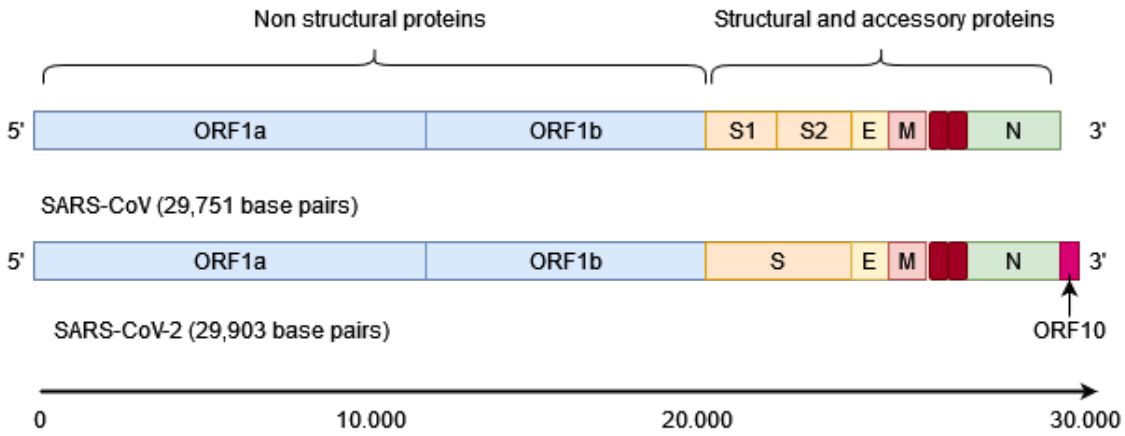


Figure 1: Genome structure of the coronaviruses SARS-CoV and SARS-CoV-2 genome. Some ORFs and accessory proteins are not drawn. ORF stands for Open Reading Frames.

Frames), ORF1a and ORF1b, collectively classified as ORF1ab, encompass two-thirds of the genome and are translated into polyproteins. These polyproteins are responsible for producing the 16 non-structural proteins, nsp1 - nsp16, which are associated with the synthesis of viral RNA. The other one-third of the genome encodes for four major structural proteins: spike (S), Membrane (M), Envelope (E), and Nucleocapsid (N) as well as some accessory proteins. The spike S glycoprotein is the largest of the structural proteins, it is made up of two subunits called S1 and S2 and plays an essential role in binding to receptors on host cells. The membrane M and envelope E proteins are collectively responsible for the regulation of virus assemblies and are the smallest among the structural proteins. Lastly, the nucleocapsid N protein protects the viral genome by packing it into a helical ribonucleocapsid. Other ORF accessory proteins are located between or inside these structural proteins, often consisting of amino acid residues and their interactions are currently not well understood. Chan, Choi, and Schork 2020 suggested that the presence of unique sequence signatures in the 3' untranslated region (3'-UTR) exists in both SARS-CoV and SARS-CoV-2 which could contribute to the survival mechanism of the viruses. Bojkova et al. 2020 also indicated that an extra ORF protein unique to SARS-CoV-2 was found in infected cells, however, there has been no conclusive evidence on the functionality of this protein.

2.3 Defining the Experiment Parameters

Now that we have an overview of the coronaviruses' genome structure, we need to define the experiment parameters:

1. The genome sequence
2. The target genome elements
3. The length of k-mers to be tested.

As previously discussed, the experiments will be conducted on the human hg38, SARS-CoV, and SARS-CoV-2 genome sequences. The target genome elements for humans are exons due to their clustered nature. For coronaviruses, despite having the largest genome size out of all RNA viruses, it is still relatively small when compared to human DNA. The human genome has a length of ∼3 billion nucleotides, approximately 100.000 times longer than coronaviruses which are roughly ∼28.000−30.000 nucleotides long. As such, we selected 2 of the most substantial genome elements, the ORF1ab and Spike S genes.

The third and last requirement is unique to the HCR algorithm that will be described in section 3.3. The shorter sequence of RNA posed an issue as the number of k-mer counts drastically reduces as the length of the k-mer increased. Using Jellyfish by Marçais and Kingsford 2011 and a python script, we compared the number of high-frequency k-mers from chromosome-1 of the human hg38 genome, and the complete genomes of SARS-CoV and SARS-CoV-2:

k-mer with frequency > 2			
k-mer length (k)	Human hg38 (chr1)	SARS-CoV (complete)	SARS-CoV-2 (complete)
2	16	16	16
3	64	64	64
4	256	256	256
5	1024	1005	988
6	4096	3121	2894
7	16384	4672	4538
8	65536	1824	2190
9	262137	290	378
10	1048568	34	41

Table 1: List of the number of valid k-mer candidates (k-mer with count > 2) for hg38, SARS-CoV, SARS-CoV-2 genomes

As expected, the k-mer frequencies for both viral genomes lack in comparison to a single chromosome from the human genome. Initial testing showed that the clustering level of some k-mers is higher than expected, indicating the experiments might still produce viable results. The size of the human genome posed a different problem, namely high computational demand, therefore we limited our testing to k-mers of length $k = 2-9$.

To sum it up, we will be running our experiments on k-mers of length $k = 2-9$. The genome elements that will be tested for in SARS-CoV and SARS-CoV-2 are the ORF1ab and Spike S genes, and the genome element that will be tested for in human hg38 will be exons.

3 Methods

In this section, we will be diving deeper into the principles and guidelines from the Hackenberg, Rueda, et al. 2012 and outline the methodology and procedure described in it. All of the code written for this project are in Python. We made use of Python's extensive bioinformatic tools such as Biopython for reading FASTA data files and Matplotlib for plotting results. Jellyfish by Marçais and Kingsford 2011 was used for fast high-frequency k-mer counting. Figure 2 shows the general workflow of the project.

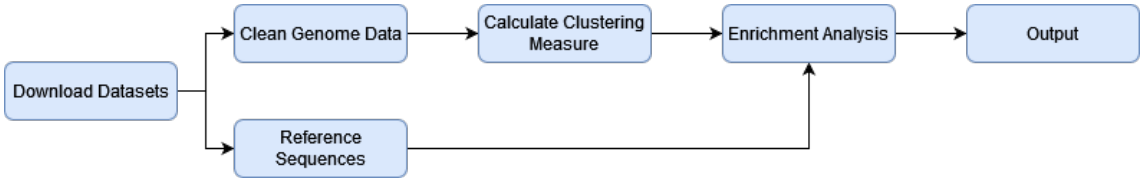


Figure 2: Workflow of the project.

3.1 Datasets

All of the genome assemblies (Human hg38, SARS-CoV, and SARS-CoV-2) and genome element reference sequences (Human exons, CoV ORF1ab gene, and CoV Spike S gene) were downloaded from the NCBI GenBank and RefSeq databases. RefSeq is a non-redundant set of reference sequences (Pruitt, Tatusova, and Maglott 2005) of genomic regions, transcripts, and proteins that have been curated. These reference sequences commonly contain the entire clusters of genome elements within them.

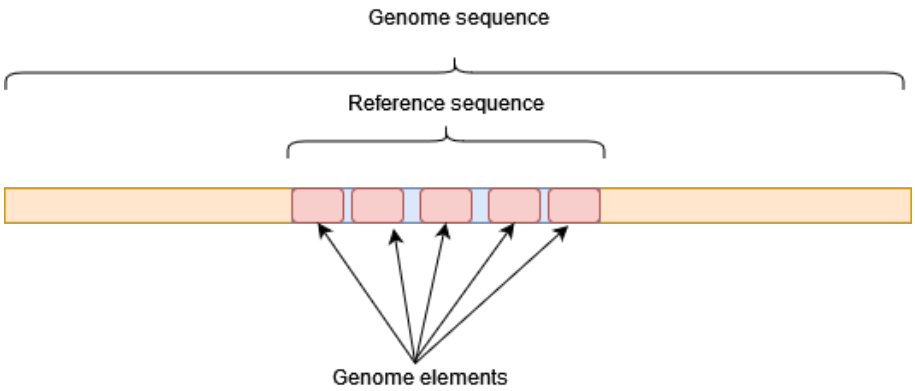


Figure 3: General overview of a genome sequence, the genome element here is a cluster of exons.

3.2 Cleaning the Genome Dataset

Since we will be measuring the clustering levels of k-mers, we need to first remove low-frequency k-mers such as k-mers with a count of 1, as a single k-mer copy cannot form a cluster by itself. Jellyfish offers a fast and efficient solution to achieve this goal. There are two ways to count high-frequency k-mers using Jellyfish, a one-pass method and a two-pass method. Both methods are based on using Bloom filters. The one-pass method provides an approximate count for some percentage of the k-mers whereas a two-pass method provides an exact count but uses more memory. For this thesis, we performed a two-pass method on the genome datasets. In the two-pass method, a Bloom counter is first created using the command:

```
jellyfish bc -m 8 -s 100M -t 16 -o sars_cov_2.bc sars_cov_2.fa
```

This creates a Bloom counter of 8-mers from the SARS-CoV-2 genome dataset. We then pass this Bloom counter into the jellyfish count command to insert only k-mers that have been seen at least twice in the first pass into the hash.

```
jellyfish count -m 8 -s 3M -t 16 --bc sars_cov_2.bc sars_cov_2.fa
```

This method of counting takes only k-mers of count greater than 1 and reduces memory usage significantly. However, the HCR algorithm requires a minimum k-mer frequency of 3. A python script was used to filter out any left-over k-mers that have a count of 2 and a Snakemake (Köster and Rahmann 2018) pipeline is built to automate the whole dataset cleaning process (see Figure 4).

Figure 4: Snakemake pipeline of the data cleaning process. On the top level - first pass of the two-pass method, the Bloom counter of each k-mer file is being created. Second level - bloom counters of the k-mer files are passed into the jellyfish count command to count only k-mers that have been seen at least twice. Third level - the jellyfish count results are dumped into readable FASTA files and processed by a python script to remove left-over k-mers in each file that has a frequency of 2. On the lowest level - the Snakemake rule produces all files required to create the final output file.

3.3 Measuring Word Clustering using HCR Algorithm

Coming back to the literary text analogy, let's assume that DNA words behave similarly to words in a text. Carpena et al. 2009, who investigated the word clustering in literary texts, mentioned that the distance distribution $P(d)$ can be used to "characterize the spatial distribution of a particular word, as well as to show the relationship between word clustering and word semantic meaning." We can directly translate this to our biological context and use $P(d)$ as a measurement of clustering. When a word is distributed randomly along the sequence, $P(d)$ follows the geometric distribution:

$$P(d) = (1-p)^{d-1}p \tag{1}$$

where d is the set of distances between consecutive words and p is the probability of occurrence. However, if the word distribution is not random, $P(d)$ will take different forms. In order to measure the clustering level of a k-mer in non-random word distributions, the normalized clustering measure σ_{nor} is used. The HCR algorithm calculates σ_{nor} of each k-mer by performing the following steps:

For each k-mer:

1. Detect all k-mer copies with a count greater than 3 in the genome sequence and store their start and end coordinates. The copies should not be overlapping each other, i.e. once a copy is found the search is resumed one index after the end of the copy.

2. Calculate the set of distances $d = (d_1, d_2, d_3, ..., d_{n-1}, d_n)$ between consecutive copies. The distance is defined as:

$$d_i = startCoordinate(d_{i+1}) - endCoordinate(d_i) \tag{2}$$

   A minimum distance of 1 can occur when two copies are located directly next to each other. Hence, when there are only 2 copies of the k-mer, it may result in false positives, where the distance between the copies is 1, thus giving an impression of high clustering but may actually be random.

3. Calculate the probability p that the k-mer occurs in the genome using the following formula:

$$p = \frac{N}{(L_s - k + 1) - N \cdot (k-1)} \tag{3}$$

   with N being the number of non-overlapping occurrences of the k-mer in the sequence, k the length of the k-mer and L_s the sequence length. The formula is simply the number of target k-mer in the sequence divided by the total number of k-mers in the sequence.

4. Calculate the standard deviation of $P(d)$ and the coefficient of variation CV:

$$\sigma = \sqrt{\langle d^2 \rangle - \langle d \rangle^2} \tag{4}$$

$$CV = \frac{\sigma}{\langle d \rangle} \tag{5}$$

$\langle d \rangle$ refers to the mean distance and CV is a form of $P(d)$ and a high CV indicates clustering. However, different k-mers of the same length can have different distance distributions and thus fluctuating CVs. Therefore it is important that these CV values are normalized.

5. Calculate the normalized clustering measure σ_{nor}:

$$\sigma_{nor} = \frac{CV}{\sqrt{1-p}} \tag{6}$$

σ_{nor} is a non-bias estimate of the clustering level of a k-mer. A $\sigma_{nor} = 1$ indicates randomness, the coefficient of variations is close or equal to the mean distance and no substantial meaning can be inferred. When $\sigma_{nor} > 1$, the coefficient of variations is greater than the mean distance, and as previously mentioned, this indicates clustering. When $\sigma_{nor} < 1$, the coefficient of variations is lesser than the mean distance, indicating a repulsion.

3.4 Enrichment Analyses

Strong word clustering is insufficient to prove that a link between a k-mer and its biological function exists. In order to establish concrete evidence, enrichment or depletion analyses are required. In enrichment/depletion analyses, the word enrichment/depletion percentages of the k-mers are used to quantify the association between the k-mers and a given genome element. This was done by measuring the enrichment ratio r_i of each k-mer and then calculating the enrichment/depletion percentages as a function of r_i and σ_{nor}.

For a given i-th k-mer, we can calculate its enrichment ratio by first defining the k-mer density within the genome as:

$$Den_i^{in} = \frac{n_i^{in}}{Len^{in}} \tag{7}$$

with n_i^{in} being the number of copies of the k-mer located within the genome element, and Len^{in} the total length of the genome element. Afterward, we define the k-mer density outside the genome in the same manner:

$$Den_i^{out} = \frac{n_i^{out}}{Len^{out}} \tag{8}$$

with n_i^{out} being the number of copies of the k-mer located outside the genome element, not overlapping a single base with the genome element, and Len^{out} the total length of the rest of the genome sequence outside of the genome element.

The enrichment ratio r_i is then measured by:

$$r_i = \frac{Den_i^{in}}{Den_i^{out}} \tag{9}$$

This ratio can then be interpreted as:

- $r_i = 1$: The k-mer can occur either inside or outside the genome element with equal probability, i.e. random.
- $r_i > 1$: The k-mer is enriched inside the genome element.
- $r_i < 1$: The k-mer is depleted inside the genome element.

The ratio r_i is an attribute of each individual k-mer and is measured independently for a given genome sequence and genome element reference sequence. For instance, the enrichment ratio for the first 6-mer "AAAAAA" of the SARS-CoV-2 genome will be different when measured within the ORF1ab and the spike S genes. Similarly, every subsequent 6-mers will have its own enrichment ratios.

We then calculate the enrichment/depletion percentage by generalizing r_i based on the normalized clustering measure σ_{nor}. First, we ordered all k-mers by its σ_{nor} value. Next, we divided them into 20 groups $\{g_1, g_2, ..., g_{20}\}$, with each group g_i having roughly an equal amount of k-mer members.

For each group, we define

- $n_{enriched}$ as the number of k-mers with $r_i \geq 2$ (twice as likely to appear inside the genome element), and
- $n_{depleted}$ as the number of k-mers with $r_i \leq 0.5$ (half as likely to appear inside the genome element)

A k-mer is then said to be highly enriched when it appears twice as frequently inside a genome element as it is outside of the genome element, i.e. the rest of the genome sequence. Likewise, a k-mer is highly depleted when it appears half as frequently inside the genome element as it is outside of it. Finally, for a given group g_i with length n_{g_i} we calculate its enrichment percentage $f_{enrichment}$ as:

$$f_{enrichment} = \frac{n_{enriched}}{n_{g_i}} \cdot 100(\%) \tag{10}$$

The depletion percentage $f_{depletion}$ can analogously be calculated using:

$$f_{depletion} = \frac{n_{depleted}}{n_{g_i}} \cdot 100(\%) \tag{11}$$

3.5 Implementation

The code implementation, as well as supplementary resources, can be found at:

https://gitlab.cs.uni-duesseldorf.de/albi/albi-students/ba-vincent-wilantara

4 Results

4.1 Word Clustering in hg38

The hg38 genome is currently the latest human reference genome that is published by NCBI and has a higher DNA quality than the hg18 genome which was used by Hackenberg, Rueda, et al. 2012 during the development of the HCR algorithm. Testing it this way has two advantages: the human genomes hg38 and hg18 differ slightly from one another due to the diversity found in some parts of the human genome. This not only gives us a unique result but also more accurate results. Moreover, the hg18 genome contained a lot of unidentified regions in the genome sequence due to low DNA quality. The higher quality of hg38 resulted in fewer of these unidentified regions, and potentially newer exon regions that hg18 was not able to uncover. This was an idea that was proposed by Lindblad-Toh et al. 2011 when discussing the limitations of hg18. For our null hypothesis, we will assume that k-mers and functional elements are not associated with each other and that no correlation between word clustering level and word enrichment percentages exists. Using the HCR algorithm, we calculated the normalized clustering measures σ_{nor} for k-mers ($k = 2-9$) of the chromosome 1 sequence of hg38:

k-mer length	N	Min σ_{nor}	Mean σ_{nor}	Max σ_{nor}
2	16	1.091	1.415	1.506
3	64	1.079	1.455	1.823
4	256	1.054	1.330	2.128
5	1024	1.029	1.278	2.635
6	4096	0.982	1.222	2.991
7	16384	0.795	1.161	4.538
8	65536	0.000	1.143	7.190
9	262137	0.000	1.125	15.378

Table 2: Statistics of the normalized clustering measures for $k = 2-9$ in hg38 (chr1) genome. N is the number of k-mers that satisfies the frequency requirement (k-mer count >2).

As shown in Table 2, all k-mers in hg38 are mostly clustered, with mean $\sigma_{nor} > 1$. The distribution of σ_{nor} values in the genome sequence shows an increasing positive skewness as the k-mer length increases. Longer k-mer sequences tend to be more distinct due to their larger sequence permutations and therefore relatively less likely to have clusters compared to shorter k-mers, which translates to lower mean σ_{nor} and min σ_{nor}. However, this distinctness inflates their max σ_{nor} values as strongly clustered long k-mers will occupy more space in the genome sequence, this tendency is very apparent on k-mers with $k > 6$. To incorporate a balanced

spectrum of short- and long-length k-mers in our test, we will use k-mers of length $k = 5-8$, this range utilizes a balanced ratio of k-mer frequency and overall clustering level.

4.2 Enrichment/Depletion Analyses of hg38

Enrichment analyses were performed on the k-mers of length $k = 5-8$ in human exons. We measured the enrichment ratio for each k-mer inside the genome element, then divided them into 20 groups and calculated their enrichment percentages (see Section 3.4). We then plotted the enrichment percentages of each k-mer group against their σ_{nor} mid values:

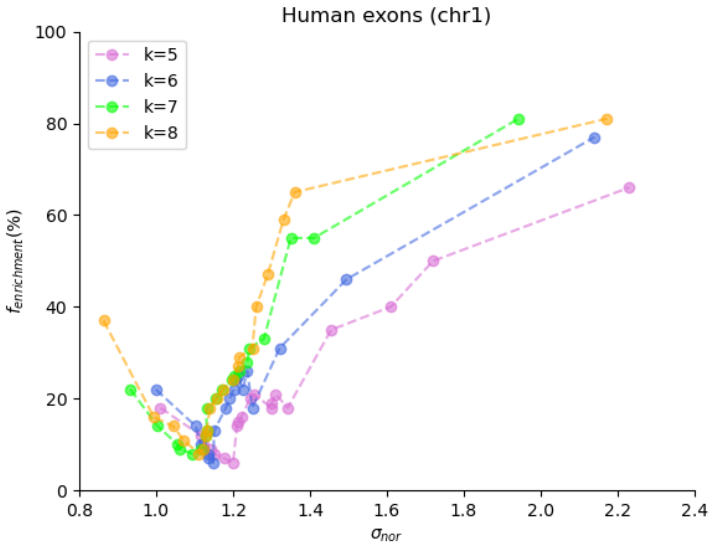


Figure 5: Enrichment percentage versus clustering level in exons. Results for 4 k-mers are shown ($k = 5-8$). In the figure, the y-axis shows the percentage of k-mers that are enriched ($r_i \geq 2$) inside the exons, and the x-axis represents the clustering level of k-mers partitioned into 20 groups of similar sizes.

We can clearly see from Figure 5 that there is a strong positive trend. To be precise, the figure indicates that k-mer groups that have higher clustering levels are more enriched inside exons and thus more likely to play a significant role in their biological function. It can then be implied that word clustering level and word enrichment percentage are strongly correlated to each other, therefore supporting the notion that k-mers are linked to functional elements of the genome. A depletion analysis is not necessary here because of the strong and apparent relationship between σ_{nor} and $f_{enrichment}$. Thus we concluded that functional k-mers can be identified in human exons by measuring the clustering level of k-mers. The top 20 clustered 6-mers in exons are listed on Table 4. We can use the result that was gained here as a litmus test for the experiments on SARS-CoV and SARS-CoV-2 genomes.

4.3 Word Clustering in SARS-CoV and SARS-CoV-2

In our research, there have been no other studies that used clustering levels to classify k-mers in the viral genomes. Given that the coronaviruses are less explored in this regard, our null hypothesis for this experiment is that no relation between k-mer clustering and biological function exists. Just as with the human genome previously, we measured the normalized clustering level σ_{nor} of each k-mer in SARS-CoV and SARS-CoV-2 using the HCR algorithm:

Genome	k-mer length	N	Min σ_{nor}	Mean σ_{nor}	Max σ_{nor}
SARS-CoV	2	16	0.943	1.029	1.239
	3	64	0.926	1.068	1.465
	4	256	0.769	1.048	1.455
	5	1005	0.000	0.937	1.980
	6	3121	0.000	0.723	1.955
	7	4672	0.000	0.444	1.858
	8	1824	0.000	0.160	1.620
	9	290	0.000	0.058	1.620
SARS-CoV-2	2	16	0.943	1.030	1.239
	3	64	0.926	1.068	1.465
	4	256	0.769	1.048	1.455
	5	988	0.043	0.959	1.980
	6	2894	0.000	0.822	1.955
	7	4538	0.000	0.670	1.857
	8	2190	0.000	0.575	1.652
	9	378	0.000	0.539	1.620

Table 3: Statistics of the normalized clustering levels of SARS-CoV and SARS-CoV-2 complete genomes. N is the number of k-mers that satisfied the frequency requirement (k-mer count >2).

Immediately, a stark contrast from the results in humans can be seen in Table 3. We noticed that the overall word clustering levels for SARS-CoV and SARS-CoV-2 k-mers are significantly lower than hg38's. This is expected due to the smaller genome size of the coronaviruses as discussed in Section 2.3. Both genomes have very close results, and we can infer from the table that most of the k-mers show signs of repulsion ($\sigma_{nor} < 1$) or no clustering at all ($\sigma_{nor} = 0/1$). Moreover, the results deviate from the pattern that we saw in Table 2. First, the minimum σ_{nor} reaches 0.000 on k-mers with much shorter lengths, this then resulted in an overall lower mean σ_{nor} on those k-mers. Second, we see that in longer k-mers ($k \geq 6$) the max σ_{nor} of the k-mers decreases unlike in the human genome, most likely due to the lower number of k-mer candidates N that meets the k-mer frequency requirement. The findings thus far support our null hypothesis.

4.4 Enrichment/Depletion Analyses of SARS-CoV and SARS-CoV-2

This section will go into detail about the outcomes of our experiments with the ORF1ab and Spike S genes in SARS-CoV and SARS-CoV-2 genomes. Enrichment analyses were first performed on the entirety of the genome elements. We measured the enrichment ratio of each k-mer of length $k = 2-9$ and then divided them into 20 groups each according to their clustering level σ_{nor}. For each k-mer group, their word enrichment percentage $f_{enrichment}$ was calculated and then plotted against its corresponding σ_{nor} mid-values. Figure 6 shows the plots for 5-mers to 8-mers:

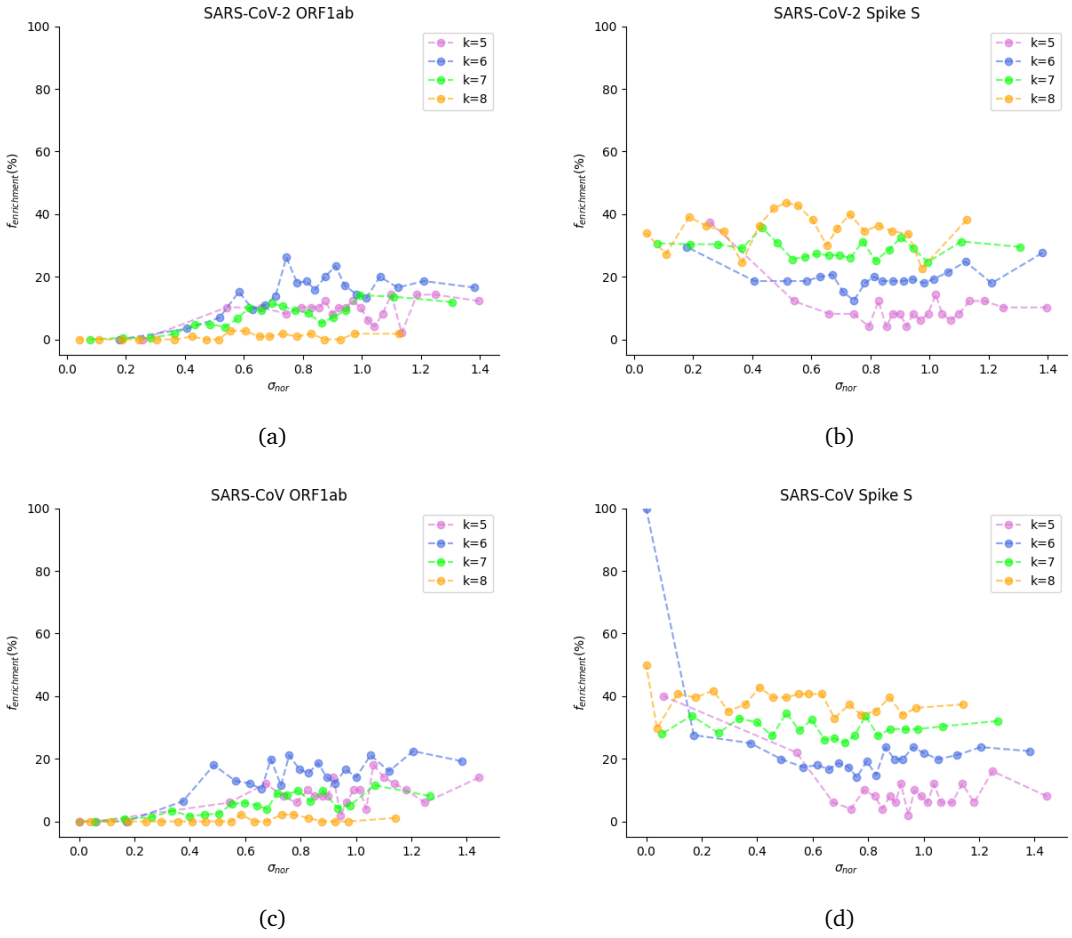
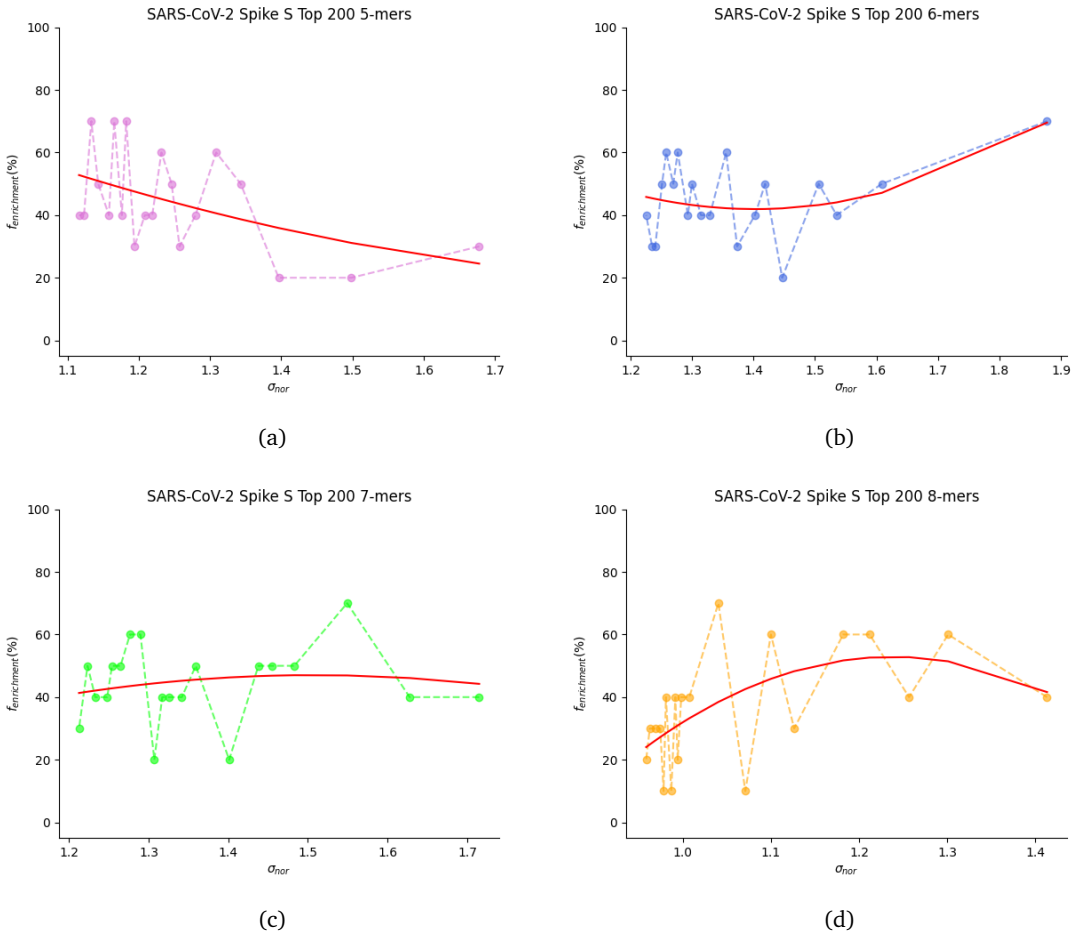


Figure 6: Plots of the word enrichment percentages against the mid-values of its corresponding word clustering levels for k-mers of length $k = 5-8$. 6a and 6b show the plot results for ORF1ab and Spike S in SARS-CoV-2 respectively. 6c and 6d show the plot results for ORF1ab and Spike S in SARS-CoV respectively. In the figure, the y-axis shows the percentage of k-mers that are enriched ($r_i \geq 2$) inside the genome element, and the x-axis represents the word clustering levels partitioned into 20 groups of similar sizes.

The plots 6a and 6c indicate an overall positive trend between word enrichment and clustering level in ORF1ab. We can see that $f_{enrichment}$ increases as σ_{nor} increases, indicating that stronger clustering equates to a higher word enrichment percentage in the genome element. On the other hand, plots 6b and 6d show no significant impact between word enrichment percentage and clustering level in Spike S. Unlike the results from the hg38 genome and ORF1ab

gene, there was no clear distinction here that directly supports or negates our null hypothesis. To investigate this further, we performed enrichment analyses on the top 200 clustered k-mers in the Spike S gene in SARS-CoV-2.



(a) (b)

(c) (d)

Figure 7: 7a, 7b, 7c, 7d shows the enrichment percentage versus the clustering level for the top 200 clustered $5, 6, 7, 8$-mers in SARS-CoV-2 respectively. Curve fitting was used to calculate the plot trend (marked by the red line). In the figure, the y-axis shows the percentage of k-mers that are enriched ($r_i \geq 2$) inside the Spike S gene, and the x-axis represents the word clustering levels partitioned into 20 groups of similar sizes.

Due to the different fluctuation of distances between the different lengths of k-mers, we plotted them separately. Using curve fitting (marked by the red line), we discovered that the top 200 clustered 6-mers (7b) have a positive trend in regards to word enrichment percentage and clustering level. In comparison, the top 200 5-mers (7a) , 7-mers (7c), and 8-mers (7d) showed neutral or negative trends. The mixed results signify that depletion analyses might be required in order to get more conclusive results. If depletion analyses indicate that the k-mers are more strongly depleted inside the genome element than they are enriched, then it is likely that functional k-mers are more enriched in other genome elements. However, if the depletion percentages are similar to the enrichment percentages, then we can conclude that no functional k-mers can be identified in the genome element using our analysis methods.

Enrichment analyses on k-mers of length $k = 2, 3, 4, 9$ were also attempted. However, each of them presented their own issues. Short-length k-mers did not respond to enrichment analyses at all, as shown in Figure 8. In the same vein, 4-mers and 9-mers are not shown here because they behaved too randomly and/or have no k-mer groups whose σ_{nor} mid-value is greater than 1.

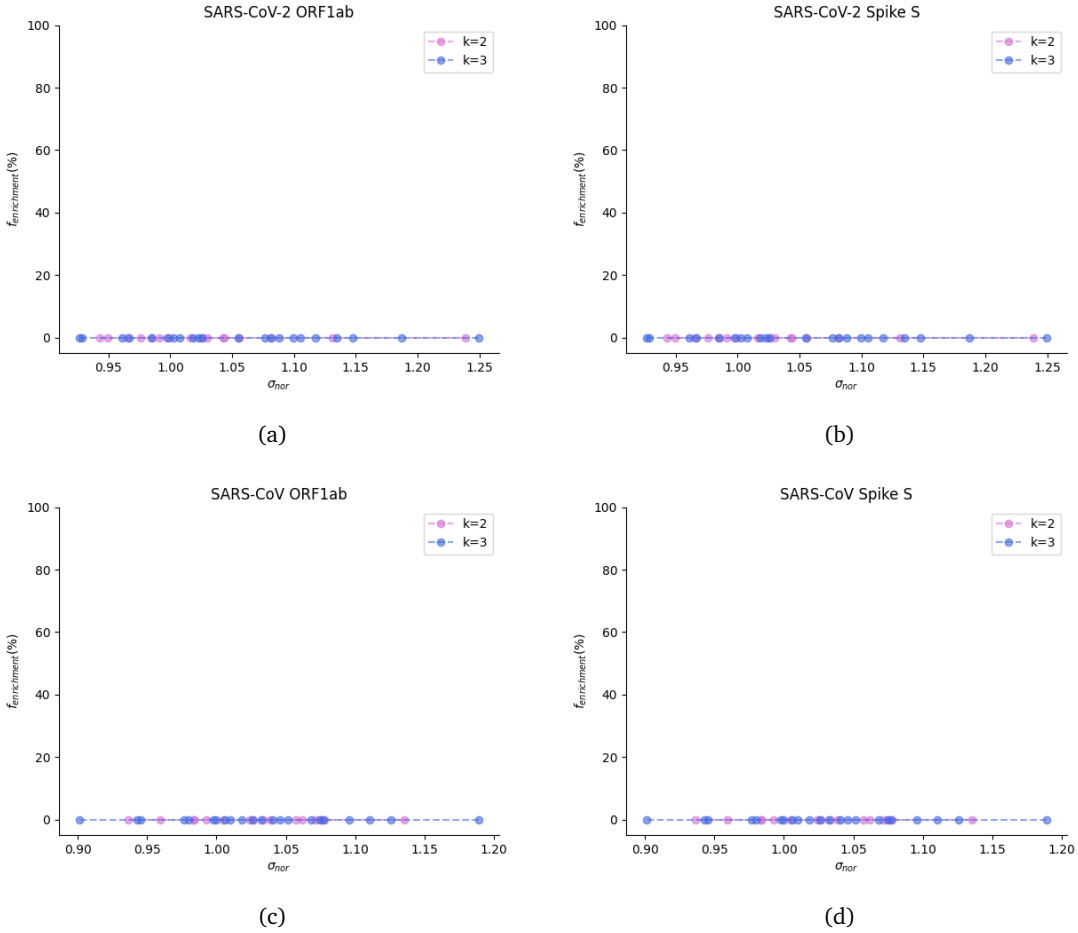


Figure 8: plots of the word enrichment percentages for 2-mers and 3-mers in the ORF1ab (8a and 8c) and Spike S (8b and 8d) gene respectively. In the figure, the y-axis shows the percentage of k-mers that are enriched ($r_i \geq 2$) inside the genome element, and the x-axis represents the word clustering levels partitioned into 20 groups of similar sizes.

We concluded that enrichment analyses were not sufficient in determining functional k-mers inside the genes. As mentioned previously, depletion analyses will dictate whether this was due to unfortunate genome element selections or the inefficacy of the clustering procedure on viral genomes. A stronger correlation between depletion percentages and clustering level corresponds to the k-mer clusters being more concentrated elsewhere in the genome, i.e. in other genome elements such as the Nucleocapsid N, Envelope E, and Membrane M genes. To confirm this we conducted depletion analyses on the ORF1ab and Spike S genes:

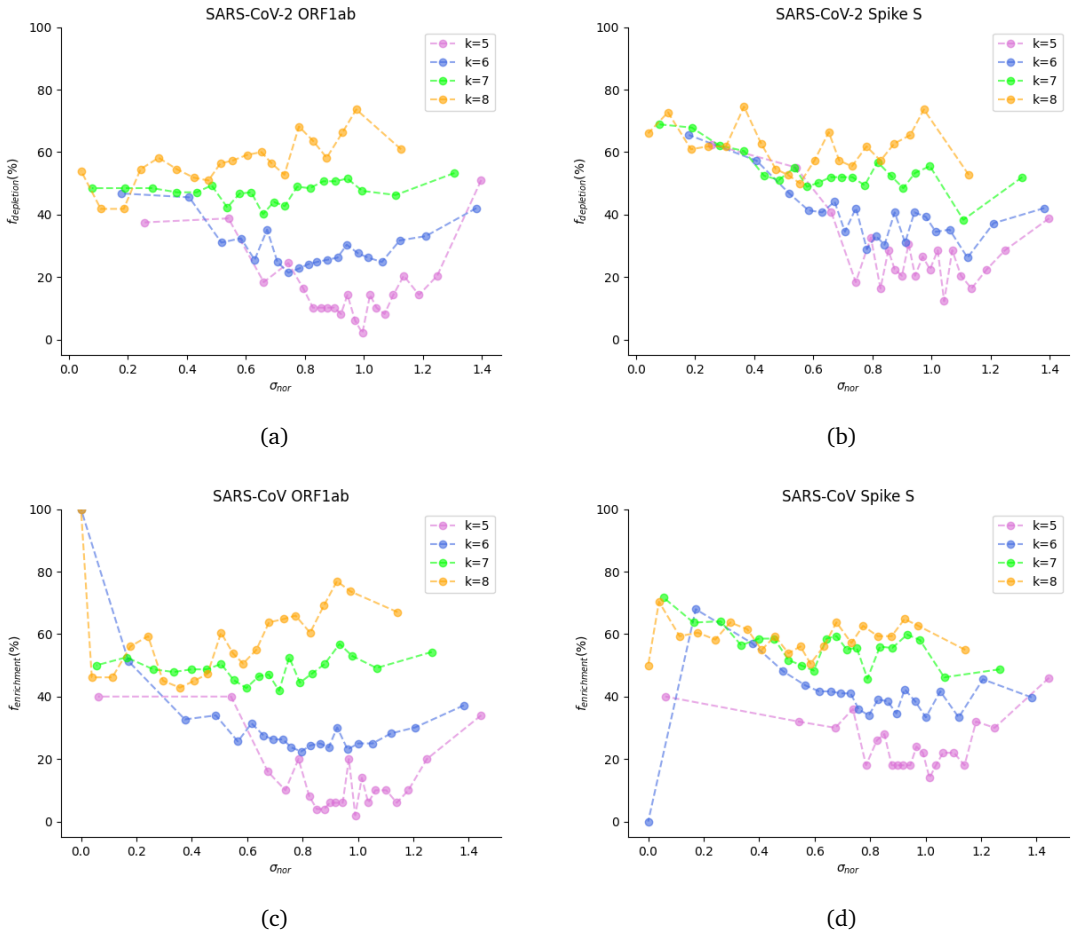


Figure 9: Plots of the word depletion percentages against the mid-values of its corresponding word clustering levels for k-mers of length $k = 5-8$. 9a and 9b show the plot results for ORF1ab and Spike S in SARS-CoV-2 respectively. 9c and 9d show the plot results for ORF1ab and Spike S in SARS-CoV respectively. In the figure, the y-axis shows the percentage of k-mers that are enriched ($r_i \geq 2$) inside the genome element, and the x-axis represents the word clustering levels partitioned into 20 groups of similar sizes.

Figure 9 shows that the word depletion percentage $f_{depletion}$ is overall greater than the word enrichment percentage found in Figure 6. In enrichment analyses, we see that the k-mers are enriched by $\sim 0-20\%$ in ORF1ab (6a, 6c) and $\sim 10-40\%$ in Spike S (6b, 6d) whereas the depletion percentages are $\sim 10-70\%$ in ORF1ab (9a, 9c) and $\sim 20-70\%$ in Spike S (9b, 9d). This verifies our assumption that the k-mers are much more depleted in the coronavirus genes than they were enriched, suggesting that the k-mers are highly enriched elsewhere in the genome.

4.5 Conclusion

In summary, we demonstrated that it is possible to quantitatively distinguish the biologically important k-mers by observing their clustering levels and enrichment percentages in human exons. Empirical data of the clustering levels indicate that k-mers in SARS-CoV and SARS-CoV-2 genomes are much less clustered compared to that in humans. Moreover, we found that k-mers are highly enriched inside human exons and highly depleted in SARS-CoV and SARS-CoV-2 ORF1ab and Spike S genes. Although no functional k-mers were identified in the SARS-CoV and SARS-CoV-2 genes, our experiments suggest that the functional k-mers are enriched in other areas of the genome. Preliminary tests on the 3' untranslated region of SARS-CoV and SARS-CoV-2 support this claim, with an average of 50% word enrichment percentage inside the genome element. This finding hopefully helps to illuminate some of the lesser-known functional regions of SARS-CoV-2.

5 Discussion

We proposed and replicated in this thesis a method of procuring functional k-mers by measuring their global clustering levels based on fluctuations in distances between consecutive k-mers and then analyzing them in genome elements. Based on our results with SARS-CoV and SARS-CoV-2, it can be argued that the HCR algorithm can be improved upon for smaller genomes such as the viral genomes. The algorithm was developed with large DNA sequences in mind, and since the experimentation using this method on the SARS-CoV-2 genome was a novel case study as far as our research has found, there were bound to be limitations to the results. One example would be false positives or false negatives, the HCR algorithm takes only k-mers of frequency > 3 into account, in large genomes, this makes perfect sense, as most k-mers will likely have more than 3 copies and any k-mer that has less than that should be regarded as random. However, in shorter genomes, the possibility of falsely filtering out potential functional k-mers is much higher, as the number of valid k-mer shrinks exponentially, as shown in Table 1. In our case, this locks our experiments from using k-mers of length $k > 9$, therefore the algorithm needs to be adapted in this regard. Alternatively, other studies have used machine learning to discern k-mer sequences (Manjunath, Prashanth, and Guru 2022), suggesting that perhaps integration with machine learning could improve upon our clustering algorithm.

Despite the low clustering levels in the coronaviruses, "conserved" k-mers were discovered between SARS-CoV and SARS-CoV-2. Hackenberg, Rueda, et al. 2012 shared the notion of "conserved words" where a set of clustered k-mers are shared amongst genomes, in order to preserve the biological functions of a genome element. Much like phylogenetic footprints, these conserved k-mers can be used to improve predictions on the functionality of genome elements. This further lends support to our hypothesis that the k-mers are enriched in other genome elements. A list of the top conserved 6-mers can be found in Table 5.

References

[1] Aaron Goldman and Laura Landweber. “What Is a Genome?” In: *PLoS genetics* 12 (July 2016), e1006181. DOI: 10.1371/journal.pgen.1006181.

[2] Heejoon Chae et al. “Comparative analysis using K-mer and K-flank patterns provides evidence for CpG island sequence evolution in mammalian genomes”. In: *Nucleic acids research* 41 (Mar. 2013). DOI: 10.1093/nar/gkt144.

[3] Aimee M. Deaton and Adrian Bird. “CpG islands and the regulation of transcription”. English. In: *Genes and Development* 25.10 (May 2011), pp. 1010–1022. ISSN: 0890-9369. DOI: 10.1101/gad.2037511.

[4] Filipe V. Jacinto and Manel Esteller. “Mutator pathways unleashed by epigenetic silencing in human cancer”. In: *Mutagenesis* 22.4 (Apr. 2007), pp. 247–253. ISSN: 0267-8357. DOI: 10.1093/mutage/gem009. eprint: https://academic.oup.com/mutage/article-pdf/22/4/247/3744066/gem009.pdf. URL: https://doi.org/10.1093/mutage/gem009.

[5] M. Ortuño et al. “Keyword detection in natural languages and DNA”. In: *EPL (Europhysics Letters)* 57 (Jan. 2007), p. 759. DOI: 10.1209/epl/i2002-00528-3.

[6] Dannie Durand and David Sankoff. “Tests for Gene Clustering”. In: *Proceedings of the Sixth Annual International Conference on Computational Biology*. RECOMB ’02. Washington, DC, USA: Association for Computing Machinery, 2002, pp. 144–154. ISBN: 1581134983. DOI: 10.1145/565196.565214. URL: https://doi.org/10.1145/565196.565214.

[7] Michael Hackenberg, Pedro Carpena, et al. “WordCluster: Detecting clusters of DNA words and genomic elements”. In: *Algorithms for molecular biology : AMB* 6 (Jan. 2011), p. 2. DOI: 10.1186/1748-7188-6-2.

[8] Michael Hackenberg, Antonio Rueda, et al. “Clustering of DNA words and biological function: A proof of principle”. In: *Journal of theoretical biology* 297 (Mar. 2012), pp. 127–36. DOI: 10.1016/j.jtbi.2011.12.024.

[9] Sarwan Ali et al. “A k-mer Based Approach for SARS-CoV-2 Variant Identification”. In: Nov. 2021, pp. 153–164. ISBN: 978-3-030-91414-1. DOI: 10.1007/978-3-030-91415-8_14.

[10] Aiping Wu et al. “Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China”. In: *Cell Host and Microbe* 27 (Feb. 2020). DOI: 10.1016/j.chom.2020.02.001.

[11] Maria Romano et al. “A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping”. In: *Cells* 9 (May 2020), p. 1267. DOI: 10.3390/cells9051267.

[12] Rimanshee Arya et al. “Structural insights into SARS-CoV-2 proteins”. In: *Journal of Molecular Biology* (Dec. 2020). DOI: 10.1016/j.jmb.2020.11.024.

[13] Agnes Chan, Yongwook Choi, and Nicholas Schork. “Conserved Genomic Terminals of SARS-CoV-2 as Coevolving Functional Elements and Potential Therapeutic Targets”. In: *mSphere* 5 (Nov. 2020). DOI: 10.1128/mSphere.00754-20.

[14] Denisa Bojkova et al. “Proteomics of SARS-CoV-2-infected host cells reveals therapy targets”. In: *Nature* 583 (July 2020), pp. 1–8. DOI: 10.1038/s41586-020-2332-7.

[15] Guillaume Marçais and Carl Kingsford. “A fast, lock-free approach for efficient parallel counting of occurrences of k-mers”. In: *Bioinformatics* 27.6 (Jan. 2011), pp. 764–770. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr011. eprint: https://academic.oup.com/bioinformatics/article-pdf/27/6/764/16902460/btr011.pdf. URL: https://doi.org/10.1093/bioinformatics/btr011.

[16] Kim Pruitt, Tatiana Tatusova, and Donna Maglott. “NCBI Reference Sequences (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins”. In: *Nucleic acids research* 33 (Feb. 2005), pp. D501–4. DOI: 10.1093/nar/gki025.

[17] Johannes Köster and Sven Rahmann. “Snakemakea scalable bioinformatics workflow engine”. In: *Bioinformatics* 34 (May 2018). DOI: 10.1093/bioinformatics/bty350.

[18] Pedro Carpena et al. “Level statistics of words: Finding keywords in literary texts and symbolic sequences”. In: *Phys. Rev. E* 79 (3 Mar. 2009), p. 035102. DOI: 10.1103/PhysRevE.79.035102. URL: https://link.aps.org/doi/10.1103/PhysRevE.79.035102.

[19] Kerstin Lindblad-Toh et al. “A high-resolution map of human evolutionary constraint using 29 mammals”. In: *Nature* 478 (Oct. 2011), pp. 476–82. DOI: 10.1038/nature10530.

[20] Ravikumar Manjunath, M. Prashanth, and Devanur Guru. “Matching Pattern in DNA Sequences Using Machine Learning Approach Based on K-Mer Function”. In: Jan. 2022, pp. 159–171. ISBN: 978-3-030-96633-1. DOI: 10.1007/978-3-030-96634-8_14.

A Appendix

Top 20 clustered 6-mers in human exons		
Rank	k-mer	σ_{nor}
1	CCCGCG	2.623409800557543
2	CGCGGC	2.55420818633698
3	CGGCGG	2.428537971056736
4	CGCGGG	2.4236820692687906
5	CCGCCG	2.416741378851082
6	CCGCGG	2.346190924187146
7	GCCGCG	2.343884955319045
8	CGCCGC	2.3403243635684734
9	GCGGCG	2.337908102240944
10	CGGCGC	2.321990067979185
11	GCGCCG	2.3183014803701005
12	CGCCGG	2.2764591382631325
13	CCGGCG	2.225221418520838
14	CGCGCG	2.2179717838918984
15	GCCCCG	2.2075372013591714
16	GCGCGC	2.1290856406303016
17	CGGGGC	2.070907159931333
18	GCGGCC	2.058350820425071
19	GGCCGC	2.058252380192431
20	CGGGCG	2.0554544111924256

Table 4: Ranked list of the top 20 clustered 6-mers with their normalized clustering level σ_{nor} in human hg38 exons.

Top conserved 6-mers in SARS-CoV and SARS-CoV-2		
k-mer	σ_{nor} (SARS-CoV)	σ_{nor} (SARS-CoV-2)
ACGAAC	1.9553092627965805	1.9599921812615817
CCAAAA	1.9234806306617496	1.2972042750896853
AACTTC	1.8944767107370326	1.2321169468423012
GTTTAA	1.8778404023614057	1.1300621940437037
TAATTA	1.8773148628322618	1.5041662512194676
AAACGA	1.632460769640049	1.7904563930818866
TTTGGC	1.6252376046649477	1.1446911493051137
AAGTGT	1.6235444441641163	1.54823214952956
AACGAA	1.6188367443941758	1.1510601604225774
ACAAGG	1.6155668053205157	1.033957577183929
TAGTCA	1.6086259352226482	1.0183375043733132
CGAACT	1.6015924200522496	1.3275451587975804
CACCAA	1.5977845339179513	1.616097715673185
AGGAAC	1.5854188634132538	1.4335106992067415
CTCTTA	1.5763783386030537	1.1801310728599583
ACCAAA	1.545334041538119	1.207577974561995
CTTGGA	1.5244729581095935	1.1042965592109963
CGAACA	1.519921667150104	1.3302702026466457
AAGGAA	1.5150164715626762	1.1883519611148547
ACCAGT	1.5145364647409019	1.0527874107574808
AGAAGT	1.5110563540609272	1.0666153897329518
AAAAAA	1.4919427676510495	1.731235388474932
ACAAAA	1.4873857995896549	1.1719799325143536
AACCAT	1.468868991543019	1.1152825406287754
ACTTCT	1.4653366155991494	1.1674004188685732
TGTACT	1.449547510269564	1.151150803844315
ATGACA	1.4466175633364278	1.3709033521595477

Table 5: List of top shared 6-mers that formed clusters in SARS-CoV and SARS-CoV-2 (sorted by the σ_{nor} value of SARS-CoV in descending order).