COVID-19 Genome Analysis using Alignment-Free Methods

M. Saqib Nawaz[1], Philippe Fournier-Viger[1], Xinzheng Niu[2], Youxi Wu[3] and Jerry Chun-Wei Lin[4]

[1]School of Humanities and Social Sciences, Harbin Institute of Technology, Shenzhen, China
[2]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
[3]Department of Computer Science and Engineering, Hebei University of Technology, Tianjin, China
[4]Department of Computing, Mathematics and Physics, Western Norway University of Applied Sciences (HVL), Bergen, Norway
msaqibnawaz@hit.edu.cn, philfv8@yahoo.com, 2386100@qq.com, wuc567@163.com, jerrylin@ieee.org

Abstract. Examining the genome sequences of the novel coronavirus (COVID-19) strains is critical to properly understand this disease and its functionalities. In bioinformatics, alignment-free (AF) sequence analysis methods offer a natural framework to investigate and understand the patterns and inherent properties of biological sequences. Thus, AF methods are used in this paper for the analysis and comparison of COVID-19 genome sequences. First, frequent patterns of nucleotide base(s) in COVID-19 genome sequences are extracted. Second, the similarity / dissimilarity between COVID-19 genome sequences are measured with different AF methods. This allows to compare sequences and evaluate the performance of various distance measures employed in AF methods. Lastly, the phylogeny for the COVID-19 genome sequences are constructed with various AF methods as well as the consensus tree that shows the level of support (agreement) among phylogenetic trees built by various AF methods. Obtained results show that AF methods can be used efficiently for the analysis of COVID-19 genome sequences.

Keywords: COVID-19, Genome sequence, Nucleotide bases, Alignment-free methods.

1 Introduction

The novel coronavirus disease (COVID-19), caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) virus, was first identified in Wuhan, China in December 2019 [1]. The World Health Organization (WHO) declared COVID-19 a pandemic on March 11, 2020 [2]. Till now, more than 132 million people have been infected by the COVID-19, with more than 2.5 million deaths worldwide. The mortality rate of SARS-CoV-2 is 3.4%, much lower than

than that of SARS-CoV-1 (9.6%) and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV (35%)) [3]. However, this disease infection rate is much greater than those of SARS-CoV-1 and MERS-CoV.

The COVID-19 genome sequence is made from single-stranded sequence of nucleotides called RNA. Identifying the sequence of neucleotides in a genome is called genome sequencing. The genome of SARS-CoV-2 has been sequenced by different groups around the world which revealed multiple strains of the virus and showed that its genome is 79% similar to the SARS-CoV-1 and 50% to the MERS-CoV, respectively [4]. COVID-19 genome sequencing and its analysis are critical to understand its behavior, its origin, how fast it mutates, and for the development of effective therapeutics or vaccines that produce long-term immunity. Thus, our focus in this paper is on the analysis and comparison of COVID-19 genome sequences with sequence alignment methods.

Sequence alignment in bioinformatics is the process of comparing and finding similarities/dissimilarities between biological sequences. The most prominent tools for sequence analysis are based on alignment-based methods (either global or local, pairwise or multiple sequence alignment (MSA). Alignment-based approaches are very popular and are generally considered the references for sequence analysis and comparison. However, they are inappropriate in some situations. For example, (1) these methods cannot obtain qualified and reliable alignment for divergent sequences. (2) These approaches are memory and time-consuming when aligning very large datasets that contain hundreds or thousands of sequences. (3) These methods are not suitable for scenarios of low sequence identity. (4) Obtained results with these methods depend on various a priori assumptions (about sequence evolution) and parameters (substitution matrices, gap opening and extension penalties, etc.) [5–8]. To overcome these limitations, various alignment-free (AF) methods [9] have been proposed.

AF methods have emerged as a natural framework in understanding the patterns and properties of biological sequences. AF methods are based on mapping symbolic sequences (that describe DNA/RNA and proteins) into vectors spaces. The main purpose of converting sequences to vectors is to apply techniques for filtering, normalization, similarity/dissimilarity calculation and clustering more efficiently. Some main advantages of using AF methods are: (1) computational inexpensiveness, (2) effortlessly dealing with whole genomes, (3) robustness to shuffling and recombination events. (4) applicability on low sequence conservation that cannot be handled by alignment. (5) no dependence on assumptions about the evolutionary trajectories of sequence changes [7, 9, 10]. AF methods are now applied to problems that range from the study of phylogenetic and regulatory elements to protein classification and sequence assembly.

In recent years, various new AF methods have been proposed. The applicability and potential of these methods for COVID-19 genome sequence analysis is an important research topic. We believe that the information that AF methods provide can not only support computational investigations of COVID-19 genome sequences but also can facilitate the clinical research. In this paper, the goal is

to analyse and compare COVID-19 genome sequences with AF methods. More specifically, various AF methods are used on COVID-19 genome sequences to:

1. Find frequent patterns of nucleotides in COVID-19 genome sequences.
2. To find similarity/dissimilarity between COVID-19 genome sequences by using different distance measures. Moreover, the performance of different distance measures are compared.
3. Investigate various AF methods for the construction of the phylogenetic tree of COVID-19 genome sequences as well as the consensus tree.

The remainder of this paper is organized as follows. Section 2 provides a background on SARS-CoV-2. Moreover, the details for AF methods and the Alfree tool that is used for COVID-19 genome analysis and comparison is also provided. Obtained results by applying AF methods on COVID-19 genome sequences are discussed in Section 3, followed by the conclusion in Section 4.

2 Analyzing COVID-19 Genome Sequences with Alignment-Free Methods

This section provides an overview of SARS-CoV-2 and AF methods that can be used for the analysis of COVID-19 genome sequences.

2.1 SARS-CoV-2

SARS-CoV-2 is a betacoronavirus with enveloped, single-stranded (positive-sense) RNA genomes of zoonotic origin [11]. The SARS-CoV-2 contains four structural proteins: (1) Spike (S), (2) Envelope (E), (3) Membrane (M) and (4) Nucleocapsid (N) (shown in Figure 1). The S, M, and E proteins make the envelope of this virus. The E protein also plays a role in the production and maturation of SARS-CoV-2. The S and M proteins are also involved in the process of virus attachment during replication. N proteins remain associated with the RNA to form a nucleocapsid inside the envelope.

SARS-CoV-2 can be contracted from animals such as bats and humans. The virus can enter the human body through its receptors, ACE2. The process of CoV entering into the host cell begins when the S protein, that comprises S1 and S2 sub-units, binds itself to the ACE2 receptor in the host cells [13]. After binding, the viral envelope fuses with the cell membrane and releases the viral genome into the target cell. The genomic material released by this virus is mRNA. In its genome range, this virus is complemented by about six to twelve open reading frames (ORFs). The genome size of the SARS-CoV-2 varies from 29.8 kb to approximately 30 kb and its genome structure follows the specific gene characteristics of known CoVs. At the 5'UTR (terminal region), more than two-thirds of the genome comprises ORF1ab that encodes ORF1ab polyproteins. Whereas at the 3'UTR, one third consists of genes that encode structural proteins (S, E, M and N), SARS-CoV-2 also contains six accessory proteins that are encoded by ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10 genes [14].

A SARS-CoV-2 genome sequence is an ordered list of nucleotides bases (Adenine-A, Guanine-G, Cytosine-C and Thymine-T). For example, Table 1

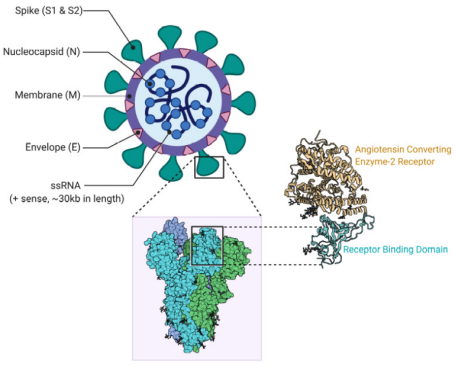


Fig. 1: SARS-CoV-2 Structure [12]

shows a sample of SARS-CoV-2 genome sequences. Note that a codon in the genome sequence represents a sequence of three nucleotide bases. There are $4^3 = 64$ different codons, in which 61 represent different amino acids that make up proteins. The remaining three codons represent the stop signals. As there are only 20 different amino acids and 61 possible codons, most amino acids are encoded by more than one codon. The genetic code defines a mapping between codons and amino acids; such that every three nucleotide bases (codon) encodes one amino acid [15]. On the other hand, *k-mers* are unique subsequences of a genome sequence of length k. For example, for $k = 1$, there are four *k-mers*: A, C, G and T. The sequence $ATCCG$ contains four $2-mers$ (AT, TC, CC and CG) and three *3-mer* (ATC, TCC and CCG).

Table 1: A sample of nucleotides in SARS-CoV-2 genome sequences

ID	Sequence
1	$\langle ....AATAACTCTATTGCCATACCCACAAATT..... \rangle$
2	$\langle ....TGCAGCAATCTTTTGTTGCAATATGGC..... \rangle$
3	$\langle ....CAGGTGCTGCATTACAAATACCATTTG..... \rangle$
4	$\langle ....CCCTAATGTGTAAAATTAATTTTAGTA..... \rangle$

2.2 Alignment-Free Methods

There are two main categories of AF sequence analysis methods:

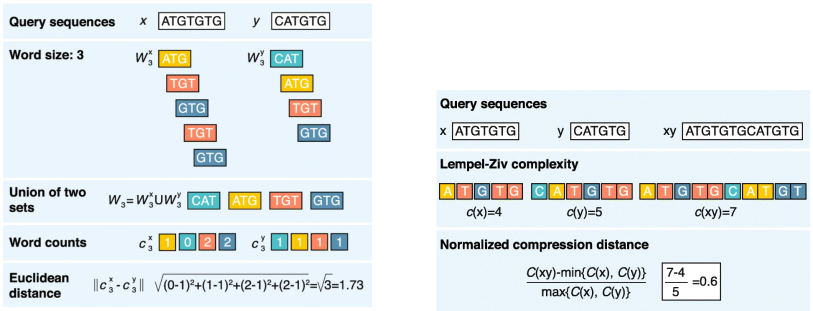
1. Word-based Methods: These methods are based on the frequencies of subsequences of a defined length, and
2. Information Theory-based Methods: These methods evaluate the informational content between full-length sequences.

Other methods, that cannot be classified in the aforementioned two groups, are based on the length of common substrings, sequence representation based on chaos theory, iterated maps, the moments of the positions of the nucleotides,

micro-alignments and Fourier transformation, etc. Interested readers can find more details on AF methods for sequence comparison in [7–10, 16–18]. Mathematically, all AF approaches are well founded in the fields of information theory, linear algebra, statistics and probability, and calculate pairwise measures of dissimilarity or distance between sequences.

Word/k-mer-based AF Methods The word/k-mer-based AF methods provide dissimilarity measures by comparing genome sequences based upon the occurrences of all *k-mers* (sequences of length k). These methods share the same working principle: similar sequences share similar words/*k-mers*, and applying mathematical operations on the occurrences of *k-mers* provide a relatively good measure for computing sequence dissimilarity. Three main steps involved in this process (shown in Figure 2(a)) are [7]:

1. Sequences division with *k-mers*: The sequences under consideration are divided into unique words of a given length (*k-mer*). Let two sequences be $x = ATGTGTG$ and $y = CATGTG$. For the word size of 3 nucleotides (*3-mer*), x and y are sliced up into: $W_3^x = \{ATG, TGT, GTG, TGT, GTG\}$ and $W_3^y = \{CAT, ATG, TGT, GTG\}$. Note that W_3^x contains three unique words (ATG, TGT, GTG). Sets of words in W_3^x and W_3^y are joined together (the union) to create a full set of words $W^3 = \{CAT, ATG, TGT, GTG\}$. The words in W^3 belong to either W_3^x or W_3^y.
2. Sequences Transformation to Vectors: After splicing, the sequences are transformed into vectors. For example, for the sequence x, its respective vector contains the number of times each particular *k-mer* (from W^3) appeared in x. Hence for x and y, the two generated vectors are: $c_3^x = (1, 0, 2, 2)$ and $c_3^y = (1, 1, 1, 1)$.



(a) Word-based (b) Information theory-based

Fig. 2: AF calculations of word-based and information theory-based distances between two sequences [7]

3. Finding dissimilarity by applying distance functions: The dissimilarity between sequences is found by applying a distance function on the sequence-representing vectors c_3^x and c_3^y. For example Euclidean distance can be used

to compute this difference [6]:

$$Eu(x,y) = \sqrt{\sum_{w \in A^k} (f_w^{(x)} - f_w^{(y)})^2}$$

where A^k represents the *k-mers* present in both sequences and $f_w^{(x)}$ and $f_w^{(y)}$ represent the frequency of the *k-mer* in x and y, respectively.
A high dissimilarity value indicates that sequences are more distant. Mapping of sequences into vectors allows one to use more than forty distance functions (such as Euclidean distance, Minkowski distance, Jaccard index, Manhattan distance, Hamming distance and Google distance) to compute the dissimilarity.

Information Theory-based AF Methods Information theory has provided successful methods for AF sequence analysis. These methods recognize and compute the amount of information shared between two sequences. As sequences made from nucleotides and amino acids are strings of symbols, so their digital organization is naturally interpretable with information theory metrics such as complexity and entropy.

For genome sequences, the Kolmogorov complexity of a sequence can be measured by the length of its shortest description. Intuitively, sequences with longer descriptions indicate a higher complexity and Kolmogorov complexity fails to find the shortest description for a given string of characters. To overcome this issue, the complexity is most commonly approximated with general compression algorithms where the length of a compressed sequence gives an estimate of its complexity This means that a more complex string will be less compressible. The process of calculating the distance between sequences using complexity (also known as compression) consists of three main steps: (shown in Figure 2(b)).

In the first step, the sequences being compared ($x = ATGTGTG$ and $y = CATGTG$) are joined to create a long sequence ($xy = ATGTGTGCATGTG$). The second step is to calculate the complexity. If x and y are exactly the same, then the complexity of xy will be very close to that of x or y. However, if x and y are dissimilar, then the complexity of xy will be close to the cumulative complexities of x and y. In the literature, many different information-based distance functions are found. For example, the Lempel–Ziv complexity [19] calculates the number of different subsequences encountered when viewing a sequence from beginning to end, as shown in Figure 2(b). In the third step, the difference between sequences is calculated by using the compressed distance measures such as normalized compression distance (NCD) [20]:

$$NCD(x,y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}}$$

where C is a compressor (for example gzip, bzip2 or PPMZ).

Another information measurement is entropy-based measures such as Kullback–Leibler divergence to compare two biological sequences. The procedure involves the calculation of the frequencies of symbols or words in a sequence and

the summation of their entropies in the compared sequences. Another measure known as Base-base correlation (BBC), a novel sequence feature, reflects genome information structure. BBC has been applied to distinguish various functional regions of genomes. A genome sequence is converted into a unique numeric vector (16-dimensional) with the following equation [17]:

$$T_{ij}(K) = \sum_{n=1}^{K} P_{ij}(n).log_2\left(\frac{P_{ij}(n)}{P_i P_j}\right)$$

where P_i and P_j are probabilities of bases i and j, $P_{ij}(n)$ is the probability of i and j at distance n in the genome and K is the maximum distance between i and j. More details on compression algorithms in AF can be found in [21].

Table 2: AF methods in Alfree

Method Name	**Distance**
Word-based methods	
Euclidean distance	d^S, d^E, d^{Eseq1}, d^{Eseq2}
Minkowski distance	$d^{Minkowski}$
Absolute -based metrics	d^{abs_mean}, d^{abs_mult}, $d^{Manhattan}$, $d^{Canberra}$
Absolute -based metrics	d^{abs_mult1}, d^{abs_mult2}, $d^{Bray-Curtis}$, $d^{Chebyshev}$
Angle metrics	d^{EVOL1}, d^{EVOL2}
Composition distance	d^{CV}
Feature Frequency Profiles	d^{FFP}
Normalized Google Distance	d^{Google}
Linear Correlation Coefficient	d^{LCC}
Return Time Distribution	d^{RTD}
Boolean vectors	$d^{Jaccard}$, $d^{Hamming}$, $d^{Sorensen-Dice}$
Frequency Chaos Game Repr.	d^{FCGR}
Information Theory-based methods	
Lempel-Ziv complexity	d^{LZ}, d^{LZ}_{*}, d^{LZ}_{1}, d^{LZ}_{*1}, d^{LZ}_{**1}
NCD	d^{NCD}
Base-Base Correlation	d^{BBC}
DNA graphical representation-Based methods	
2D Graph DNA Representation	d^{2DSV}, d^{2DMV}, d^{2DNV}

Alfree [7][1] is a pairwise and multiple AF sequence comparison tool with a web support. Methods in Alfree calculate distances between sequences by discovering various patterns and properties in unaligned sequences. It implements 38 popular AF methods to calculate distances among given nucleotide or protein sequences, tree construction and creating consensus trees. The consensus phylogenetic tree provides an estimate for the level of support (agreement) between various individual methods' trees. This allow examining the reliability of given phylogenetic relationships among different methods. Table 2 lists the word-based and information theory-based AF methods implemented in Alfree. Besides these methods, Alfree provides implementation for some graphical representation-based methods, such as d^{2DSV}, d^{2DMV} and d^{2DNV}. In such methods, features from the

[1] www.combio.pl/alfree

graphical representation of DNA sequences are used to find the essence of the base composition and distribution of sequences in a quantitative manner.

3 Experiments and Results

This section presents results obtained by applying the AF methods discussed in the previous section on the COVID-19 genome sequences. The online genome sequence database GenBank [22] was used to acquire sequencing data for strains of SARS-CoV-2. It is maintained by the National Center for Biotechnology Information (NCBI) and is built primarily by submissions from individual laboratories and large-scale sequencing centers. Statistics about the collected genome sequences are presented in Table 3, where ID is the accession number of the genome sequence. The NCBI GenBank offers to download each sequence in the form of nucleotide, coding region or protein. We downloaded the genome sequences in nucleotide form. The first genome sequence for COVID-19 (NC_045512) in Table 3 is the RefSeq in NCBI, as this was the first genome sequence released by Shanghai Public Health Clinical Center & School of Public Health [1].

Table 3: Characteristics of COVID-19 genome taken from NCBI

ID	Release Date	Length	Location	Collection Date
NC_045512	2020-01-13	29903	China	2019-12
MW052550	2020-11-03	719	South Korea	2020-07-07
MW192918	2020-10-31	654	Gabon	2020-03-14
MW173089	2020-10-26	3819	USA	2020-04-25
MW165491	2020-10-24	3822	Iran	2020-04
MW161041	2020-10-23	4043	Russia	2020-06-04
MW092768	2020-10-12	2383	Sweden	2020-02-25
MW040503	2020-09-26	1009	Venezuela	2020-05-22
MT843234	2020-08-28	287	Italy	2019-12-18
MT750057	2020-07-13	29782	USA	2020-06-17
MT750058	2020-07-13	29782	USA	2020-06-09
MT291827	2020-04-06	29858	China	2019-12-30
MT291828	2020-04-06	29858	China	2019-12-30

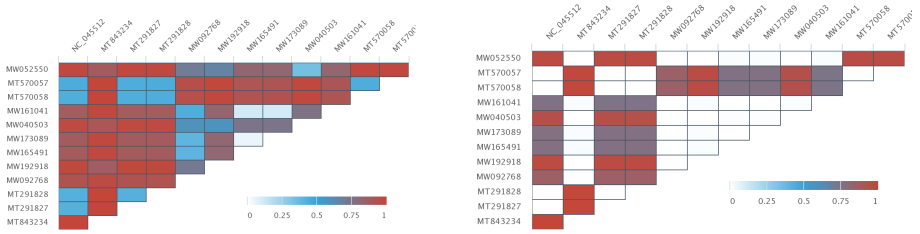
We first run Alfree to extract frequent words (nucleotide sets) from the corpus that contains sequences listed in Table 3. Obtained results are shown in Table 4, that lists the extracted frequent words in one genome sequences (NC_045512) and also in all the sequences. The first four extracted patterns (of length 1) show the total occurrence of nucleotides in genome sequences. For NC_045512, two nulceotides A and T make up for 62% of the sequence (approximately 30% for A and 32% for T). Moreover, C and G makes up the remaining 38% of the sequence (approximately 19.6% for G and 18.4% for C). For all genome sequences in the corpus, the content of A and T is 61.3% (approximately 29.9% for A and 32.2% for T), and 37.9% for C and G (approximately 19.5% for G and 18.4% for C). The extracted frequent codons (sequences of 3 nucleotides) and frequent pattern of length six (two codons) are also listed in Table 4. Interestingly, the extracted frequent words in one genome and all genome sequences are almost the same.

Next, the pairwise dissimilarities among the COVID-19 genome sequences is calculated through various distance measure methods in Alfree. The K is set

Table 4: Extracted frequent words of nucleotides

Patterns	**Occurrence**	**Patterns**	**Occurrence**
NC_045512		**All sequences**	
T	9594	T	53442
A	8954	A	49532
G	5863	G	32358
C	5492	C	30587
TTT	1004	TTT	5701
AAA	923	AAA	5009
TTA	876	TTA	4846
TGT	858	TGT	4734
TTG	817	TTG	4586
ACA	809	ACA	4509
TTGTTA	42	TTGTTA	231
TGTTAA	41	TGTTAA	223
GGTGTT	35	GGTGTT	210
TTTTAA	35	TTGGTG	206
TGTTGT	34	TTTTAA	202
TGTTGT	34	CTTTTG	196

to 3 for all word-based distance measures. Figure 3 shows the heatmap for the normalized compression distance (d^{NCD}) and standard Euclidean (d^{E}) distance measures. We found that the two methods produced two different heatmaps as d^{E} is a word-based method and d^{NCD} is based on information theory. Due to the space limitation, heatmaps for other methods are not presented here.



(a) d^{NCD} (b) d^{E}

Fig. 3: Distance between genome sequences calculated using d^{NCD} and d^{E}

Table 5 compares the performance of various word-based and information theory-based AF methods for two sequences (MT750057 and MT750058). The calculated dissimilarity by two Euclidean distance-based measures (d^{S} and d^{E}) was very different, whereas d^{Eseq1} and d^{Eseq2} yield the same results. $d^{Minkowski}$ and d^{S} indicated same similarity and the same is observed for Angle metrics measures (d^{EVOL1} and d^{EVOL2}). The performance of absolute-based metrics measures (d^{abs_mult}, d^{abs_mult1} and d^{abs_mult2}) was same. Whereas $d^{Manhattan}$ performed differently than other absolute-based metrics-based measures. On the

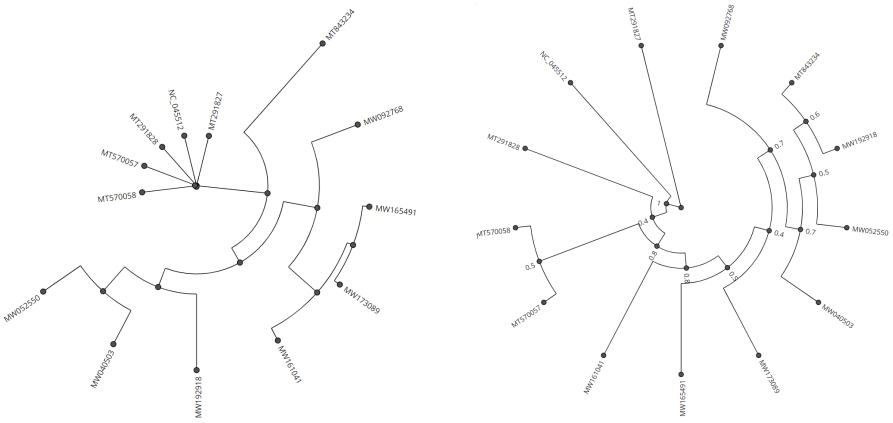
other hand, the performance of three Boolean vectors-based measures ($d^{Jaccard}$, $d^{Sorenson\_dice}$ and $d^{Hamming}$) was same. The D^{KL} is the Kullback–Leibler divergence measure that uses entropy-based measure for genome sequences comparison. Overall, it was observed that one can easily and efficiently compare the performance of different AF methods on COVID-19 genome sequences in Alfree to get useful insights about this virus.

Table 5: Distance measure values for two sequences

Measures	**AD***	**ND****	**Measure**	**AD**	**ND**
MT570057, MT570058					
d^{E}	68	0	d^{S}	8.246	0.002
d^{Eseq1}	0.0022	0.000	d^{Eseq2}	0.0022	0.000
$d^{Minkowski}$	8.246	0.002	d^{abs_mean}	0.625	0.001
d^{abs_mult}	1.219	0.003	d^{abs_mult1}	1.097	0.003
d^{abs_mult2}	1.219	0.003	$d^{Manhattan}$	40	0.001
d^{Bray_Curtis}	0.0006	0.001	$d^{Canberra}$	0.0564	0.001
d^{EVOL1}	9.798	0.000	d^{EVOL2}	9.798	0.000
d^{FFP}	6.486	0.000	d^{Google}	0.0006	0.001
d^{LCC}	0.000	0.000	d^{FCGR}	8.246	0.002
d^{KL}	2.890	0.000	$d^{Sorenson\_Dice}$	0	0.000
$d^{Jaccard}$	0.00	0.000	$d^{Hamming}$	0.000	0.000
d^{RTD}	0.001	0.002	d^{CV}	0.0001	0.000
d^{NCD}	0.438	0.443	d^{BBC}	0.001	0.003
d^{2DSV}	2.159	0.000	d^{2DMV}	2.195	0.000
d^{2DNV}	1.944	0.000	$d^{Chebyshev}$	4	0.001

*AD = Actual distance, **ND = Normalized distance

Lastly, the phylogenetic tree for COVID19 genome sequences are shown in Figure 4. A phylogenetic or evolutionary tree (also known as phylogeny) diagrammatically describes the evolutionary history and relationship of an organism or group of organisms. Phylogenetic relationships provide valuable information on shared ancestry but not necessarily on how organisms are similar or different. Figure 4(a) shows the pyhlogenetic tree for d^{NCD} as a phylogram and the phylogram in Figure 4(b) represents a majority-rule consensus tree that summarizes the agreement among various AF methods. This tree is constructed for 27 AF methods (22 word-based methods, 2 information theory-based methods and 3 DNA graphical representation-based methods). For every node in the consensus tree, the support values is represented in the range [0, 1]. These trees describe how different strains are connected with other and how they have evolved over time. A recent study [23] used an AF method, called the Natural Vector method, to analyze the phylogeny of SARS-CoV-2 with human coronaviruses. In this paper, we observe that AF methods can also analyse and compare genome sequences efficiently. Moreover, Alfree also provides the features for extracting frequent patterns for nucleotides as well as the consensus phylogenetic tree.



(a) d^{NCD} (b) Consensus tree

Fig. 4: Phylogeny for COVID-19 strains

4 Conclusion

The genome sequences for COVID-19 strains, taken from NCBI's GenBank, are compared and analyzed with AF methods for: (1) extracting frequent patterns of nucleotides in COVID-19 genome sequences, (2) finding the similarity/dissimilarity between COVID-19 genome sequences by using different distance measures and their performance comparison and (3) Phylogeny construction with various AF methods for COVID-19 genome sequences. Our experiments and obtained results show that AF methods provide an efficient framework for the analysis of COVID-19 genome sequences to get useful insights about this virus. As discussed in [24], including more COVID-19 data and applying more algorithms will allow researchers to obtain results that may guide the clinical research for this pandemic.

In the future, we plan to analyse protein sequences for COVID-19 strains with AF methods. Another direction is to use other AF similar tools like CAFE [6] for more analysis and comparison. This will also enable us to compare the aforementioned tools with each other. We are also interested in analyzing public reactions to the COVID-19 pandemic [25].

References

1. F. Wu et al. New coronavirus associated with human respiratory disease in China. *Nature*, 579:265–269, 2020 .
2. D. Cucinotta and M. Vanelli. WHO declares COVID-19 a pandemic. *Acta Biomedica*, 91(1):157–160, 2020.
3. S. Perlman. Another decade, another coronavirus. *New England Journal of Medicine*, 382(8):760–762, 2020.

4. R. Lu et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *The Lancet*, 395(10224):565–574, 2020.
5. Y. Kang et al. PVTree: A sequential pattern mining method for alignment independent phylogeny reconstruction. *Genes*, 10(2):73, 2019
6. Y. Y. Lu et al. CAFE: aCcelerated Alignment-FrEe sequence analysis. *Nucleic Acids Research*, 45(Web Server issue):W554–W559, 2017.
7. A. Zielezinski et al. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology*, 18, 186, 2017.
8. S. Vinga. Information theory applications for biological sequence analysis. *Briefings in Bioninformatics*, 15(3):376–389, 2014.
9. S. Vinga and J. Almeida. Alignment-free sequence comparison— A review. *Bioinformatics*, 19:513–523, 2003.
10. A. Zielezinski et al. Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, 20, 144, 2019.
11. M. S. Nawaz et al. Using artificial intelligence techniques for COVID-19 genome analysis. *Applied Intelligence*, https://doi.org/10.1007/s10489-021-02193-w, 2021.
12. M. Cascella et al. Features, evaluation, and treatment of coronavirus. In: StatPearls [Internet], NBK554776, https://www.ncbi.nlm.nih.gov/books/NBK554776/
13. H. Xu et al. High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa. *International Journal of Oral Research*, 12(8), 2019
14. R. A. Khailany, M. Safdar and M. Ozaslanc. Genomic characterization of a novel SARS-CoV-2. *Gene Reports*, 19:100682, 2020.
15. J.-J. Shu. A new integrated symmetrical table for genetic codes. *Biosystems*, 151:21–26, 2017.
16. J. Ren et al. Alignment free sequence analysis and applications. *Annual Review of Biomedical Data Science*, 1:93–114, 2018.
17. O. Bonham-Carter et al. Alignment-free genetic sequence comparisons: A review of recent approaches by word analysis. *Briefings in Bioinformatics*, 15(6):890–905, 2014.
18. J. Song et al. New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing. *Briefings in Bioinformatics*, 15(3):343–353, 2014.
19. H. H. Otu and K. A. Sayood. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, 19(1):2122–2130, 2003
20. M. Li et al. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–64, 2004
21. R. Giancarlo, S. E. Rombo and F. Utro. Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies. *Briefings in Bioinformatics*, 15(3):390–406, 2014.
22. E. W. Sayers et al. Genbank. *Nucleic Acids Research*, 48(D1):D84–D86, 2019
23. R. Dong et al. Analysis of the hosts and transmission paths of SARS-CoV-2 in the COVID-19 outbreak. *Genes*, 11(6): 637, 2020.
24. M. A. Ahsan et al. Bioinformatics resources facilitate understanding and harnessing clinical research of SARS-CoV-2. *Briefings in Bioinformatics*, bbaa416, 2020.
25. S. Noor et al. Analysis of Public Reaction to the Novel Coronavirus (COVID-19) Outbreak on Twitter. *Kybernetes*, Emerald Publishing, DOI: 10.1108/K-05-2020-0258