AUTOMATIC IDENTIFICATION OF SARS CORONAVIRUS USING COMPRESSION-COMPLEXITY MEASURES

A PREPRINT

Karthi Balasubramanian
Department of Electronics and Communication Engineering
Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidyapeetham, INDIA.
b_karthi@cb.amrita.edu

Nithin Nagaraj
Consciousness Studies Programme
National Institute of Advanced Studies
Indian Institute of Science Campus
Bengaluru, Karnataka, INDIA.
nithin@nias.res.in

March 24, 2020

Note: The main ideas and results of this research were first presented at the *International Conference on Nonlinear Systems and Dynamics* (**CNSD-2013**) held at Indian Institute of Technology, Indore, December 12, 2013. In this manuscript, we have extended our preliminary analysis to include SARS-CoV-2 virus as well.

ABSTRACT

Finding vaccine or specific antiviral treatment for global pandemic of virus diseases (such as the ongoing COVID-19) requires rapid analysis, annotation and evaluation of metagenomic libraries to enable a quick and efficient screening of nucleotide sequences. Traditional sequence alignment methods are not suitable and there is a need for fast alignment-free techniques for sequence analysis. Information theory and data compression algorithms provide a rich set of mathematical and computational tools to capture essential patterns in biological sequences. In 2013, our research group (Nagaraj et al., Eur. Phys. J. Special Topics 222(3-4), 2013) has proposed a novel measure known as Effort-To-Compress (ETC) based on the notion of compression-complexity to capture the information content of sequences. In this study, we propose a compression-complexity based distance measure for automatic identification of SARS coronavirus strains from a set of viruses using only short fragments of nucleotide sequences. We also demonstrate that our proposed method can correctly distinguish SARS-CoV-2 from SARS-CoV-1 viruses by analyzing very short segments of nucleotide sequences. This work could be extended further to enable medical practitioners in automatically identifying and characterizing SARS coronavirus strain in a fast and efficient fashion using short and/or incomplete segments of nucleotide sequences. Potentially, the need for sequence assembly can be circumvented.

Keywords SARS coronavirus · SARS-CoV-1 · SARS-CoV-2 · identification · COVID-19 · compression complexity · Lempel-Ziv · Effort-To-Compress · data compression

1 Introduction

SARS (Severe Acute Respiratory Syndrome) is a viral respiratory disease caused by the SARS coronavirus (SARS-CoV[1]) and having flu-like symptoms. It was first identified in Guandong province, China, in 2002 and spread rapidly to different parts of the world in a span of just a few months [1]. The primary route of transmission of the SARS coronavirus is through mucosal contact with respiratory droplets or fomites of infected persons. Marra *et al.* [1] and Rota *et al.* [2] have done extensive studies to show that SARS-CoVs forms a separate group of coronaviruses and are not closely related to other previously sequenced coronaviruses (mammalian and avian viruses).

[1] SARS-CoV-1 and SARS-CoV-2.

SARS-CoV-2 is the latest strain of the coronavirus, first discovered in late 2019 in Wuhan, China, that is responsible for the ongoing pandemic of coronavirus disease 2019 (COVID-19). Apart from SARS-CoV-1 and SARS-CoV-2, there are hundreds of other strains of SARSr-CoV (Severe Acute Respiratory Syndrome-related coronavirus) that are known to infect only non-human species such as bats (and palm civets and other mammals). SARS-CoV-2 is highly contagious in humans with the World Health Organization (WHO) designating it as a pandemic, the first ever caused by a coronavirus.

Pandemics such as the ongoing COVID-19 virus that leads to enormous loss of life globally can only be controlled by finding a vaccine or a very effective antiviral treatment. Finding a vaccine or specific antiviral treatment for such global pandemic of virus diseases requires rapid analysis, annotation and evaluation of metagenomic libraries to enable quick and efficient screening of nucleotide sequences. Traditional sequence alignment methods are not suitable since they are computationally intensive and cannot be easily scaled up as the number of sequences increase. Thus, there is a need for fast alignment-free techniques for sequence analysis [3, 4]. Further, one may have only short segments and/or incomplete fragments of nucleotide sequences to analyze [5]. Information theory and data compression algorithms provide a rich set of mathematical and algorithmic/computational tools to capture essential patterns in data that could be used for matching nucleotide sequences.

Genome sequences are inherently described by character strings and hence amenable to mathematical and computational techniques for extracting information. Exactly what information is being sought from such character strings depends on the string itself and the domain as well as the kind of application. Some targets of interest for analyzing genome sequences include:

- Various genes that constitute the genome.
- Identifying the origin of the genome sequence.
- Understanding the information content present in the coding and non-coding regions.
- Reconstructing the phylogenetic tree to study evolutionary patterns.

An important objective is to automate the above tasks so that a large number of sequences can be quickly, robustly and efficiently analyzed (as one of the steps in the endeavor for finding a vaccine).

A cursory glance at these character strings doesn't tell us much about how they can be used for these applications. But a harmonious blending of complexity analysis with the field of information theory provides deep insight in this regard. Application of complexity measures on these information bearing character strings may reveal many surprising features that generally can't be discerned by intuition or visual inspection of the data alone.

In this study, we propose a novel compression-complexity based distance measure for sequence analysis and identification. The paper is organized as follows. In section 2, a brief overview of genetic sequences and methods of analysis are described. Section 3 deals with the materials used (genome primary sequence data with their details) and the methods proposed in this study (novel distance measure for identification). Results on real data (nucleotide sequences) followed by their analysis and discussion can be found in section 4. We conclude with future research directions in section 5.

2 Genetic Sequences and their Analysis: An Overview

The basic building blocks of DNA and RNA are primary nucleobases, namely Cytosine (C), Guanine (G), Adenine (A), Thymine (T) and Uracil (U). A, C, G and T occur in DNA sequences and are known as DNA bases while A, C, G and U occur in RNA sequences and are called RNA bases. A string of these nucleobases forms a nucleic acid sequence that has the capacity to represent information. These information strings are called genetic sequences. Each species has unique characteristics differentiating it from other species and these characteristics are defined by the information content of the DNA sequences [6].

2.1 Genome and gene

The total DNA content (RNA for viruses) of an organism is known as the genome, thus representing the entire information coded in a cell, while a gene represents a section of the DNA that codes for RNA or protein. A genome consists of a sequence of multiple genes interspersed with non-coding sequences of nucleic bases [6, 7].

2.2 Genome sequence comparison

Genome data classification comes under the broad field of bioinformatics, an established multidisciplinary field for over three decades, encompassing physical and life sciences, computer science and engineering. Many fundamental problems in the fields of medicine and biology are being tackled using the tools of bioinformatics. The main requirement for

accomplishing such tasks is the availability of sequenced genome data. This has been the focus of researchers for the past few decades and efforts have been put by the National Institutes of Health (NIH) to establish Genbank®[2], a genetic sequence database containing annotated collection of all publicly available DNA sequences. Ever since its inception in 1982, there has been an exponential rise in the number of sequences in Genbank. This has provided the required resources for researchers and industry people alike for delving in to the field of bioinformatics.

Among the various aspects involved in bioinformatics, one key element is sequence comparison or analysis of sequence similarity [8, 9]. This is used in database searching, sequence identification and classification, phylogenetic tree[3] creation and in gene annotation and evolutionary modeling. Since it is impossible to recreate/simulate past evolutionary events, computational and statistical methods for comparison of nucleotide and protein sequences are used for these kinds of studies [10, 11].

There are basically two kinds of sequence comparison methods:

- Alignment based methods: These involve either shifting or insertion of gaps in sequences for alignment of two or more sequences, which make these methods computationally intensive.
- Alignment-free methods: These are computationally less intensive methods that consider the genome sequences as character strings and use distance-based methods involving frequency and distribution of bases [12–14]. Our focus in this paper is on alignment-free methodology, especially on using complexity measures for sequence comparisons.

Sequence comparison and genome data classification got a boost in the early 1990s with the use of data compression algorithms that have the ability to identify regularities in sequences [15]. They provided a means to define distances between two sequences that greatly aided in the comparison of sequences. The history behind the usage of data compression algorithms in this field has been elucidated by Otu *et al.* in [15]. We succinctly summarize that history here.

The first attempt at using data compression for phylogenetic tree construction was by Grumbach *et al.* in [16]. They explored the idea of compressing a sequence S using a sequence Q, where the degree of compression obtained by doing so would be an indicator of the distance between them. Although their definition was not mathematically valid, it set a platform for researchers to explore in this area. Varre *et al.* [17] defined a transformation distance when sequence Q is transformed to sequence S by various mutations like segment-copy, segment-reverse copy and segment-insertion. Li *et al.* [18] define a relative distance measure by using a compression algorithm called GenCompress [19] that is based on approximate repeats in DNA sequences. Using the concept of Kolmogorov complexity, the compression algorithm has been used to propose a distance between sequences S and Q. But Kolmogorov complexity, [20] being an algorithmic measure of information and a theoretical limit, can't be directly computed but only approximately estimated [21]. Hence it is not an optimum choice as a complexity measure.

Even though the idea of relative distance is an efficient one, GenCompress is a complicated algorithm that is computationally intensive. To overcome the above mentioned difficulties, Otu *et al.* [15] proposed similar but computationally simpler relative distance measures based on the Lempel-Ziv (LZ) [22] complexity measure. Given two sequences S and Q, sequences SQ and QS are formed by concatenation[4]. These four sequences are used to define four distance measures using the LZ complexity measure, as given below:

$$d(S,Q) = max\{LZ(SQ) - LZ(S),\ LZ(QS) - LZ(Q)\}. \tag{1}$$

To eliminate the effect of length of the sequence, a normalized measure is defined as follows:

$$d(S,Q) = \frac{max\{LZ(SQ) - LZ(S),\ LZ(QS) - LZ(Q)\}}{max\{LZ(S), LZ(Q)\}}. \tag{2}$$

A third distance metric based on *sum distance* is defined as follows:

$$d(S,Q) = LZ(SQ) - LZ(S) + LZ(QS) - LZ(Q). \tag{3}$$

Finally, normalized version of the sum distance is defined as:

$$d(S,Q) = \frac{LZ(SQ) - LZ(S) + LZ(QS) - LZ(Q)}{LZ(SQ)}. \tag{4}$$

[2] http://www.ncbi.nlm.nih.gov/genbank

[3] A phylogenetic tree, also called an evolutionary tree, is a tree diagram that shows the evolutionary relationships among different species according to the composition of their genes.

[4] Q is appended at the end of sequence S to yield the new sequence SQ.

Using these distance measures on mtDNA (mitochondrial DNA) samples of a wide range of eutherans (placental mammals), they have successfully re-created phylogenetic trees showing the evolutionary patterns. Other researchers have used these and slight variants of these measures to identify families of coronaviruses, mammals, vertebrates and salmons. Interested readers are referred to [23–30] for further details on these. Apart from these complexity based measures, distance measures using Markov chain models [31–33] and measures of probability [34–36] have also been proposed for the study of genome identification.

2.3 Effect of data length on complexity of sequences

Monge *et al.* in [29] have pointed out that complexity is not uniform throughout a genome. Regions including genes are more regular and have less complexity than regions that don't include a gene. This raises the issue of the length of the genome to be analyzed. Since the complexity is not uniform, it will be inaccurate to use the entire genome sequence for analysis and may possibly give erroneous results. Also the use of complete genome/gene sequences is computationally intensive and is practically infeasible for scaling up for matching a large number of genetic sequences. [5].

Our primary interest in this work lies in showing that it is not necessary to have the entire genome/gene for complexity analysis. In this work, we use LZ and Effort-To-Compress (ETC) [37] complexity measures to analyze short length segments that are randomly chosen from genome sequences. In particular, short length contiguous segments (length < 100) are randomly chosen from the sequence for analysis.

3 Materials and Methods

In this section, we describe in detail the data that was used in this study as well as the various methods that were employed for automatic identification of sequences.

3.1 Sequence analysis of coronaviruses (SARS-CoV-1)

We first analyzed the genome primary sequences of the following viruses: SARS coronavirus Urbani (AY278741.1), SARS coronavirus BJ01 (AY278488.2) and Avian Infectious Bronchitis coronavirus reference sequence (NC_001451.1). The first two virsues belong to the SARS-CoV-1 strain. The genome sequences were obtained from Genbank database, the details of which were mentioned in Section 2.2. Table 1 gives the details of the genome sequences that we use for the analysis.

Table 1: Genbank accession number, name, abbreviation and length of coronaviruses used for analysis.

S.No	Accession Number	Genome	Abbreviation	Length
1	AY278741.1	SARS coronavirus Urbani	Urbani	29727
2	AY278488.2	SARS coronavirus BJ01	BJ01	29725
3	NC_001451.1	Avian infectious bronchitis coronavirus reference sequence	IBV	27608

3.1.1 0-1 sequences

The first step in our analysis is the conversion of primary sequences into 0-1 sequences before we can evaluate complexity values. For this, we have considered three different methods to categorize the nucleotide bases and map them to the symbols 0 and 1 based on their:

- Chemical structure
- Carboxylic acid group they belong to
- Strength of the hydrogen bonds

The four bases are mapped into two classes (labelled 0 and 1) in order to create a 0-1 sequence. For every input primary sequence, we create three independent sets of 0-1 sequences using three different methods as described in Table 2. It has been shown in [38] that these three characteristic sequences give the complete information of the primary sequence.

Table 2: Mapping of DNA bases in to three independent sets of 0-1 sequences using three different methods.

Set No.	Method based on	Bases mapped to 0	Bases mapped to 1
1	Chemical Structure	Purine- {A,G}	Pyrimidine {C, T}
2	Carboxylic acid group	Amino- {A,C}	Keto-{G,T}
3	Strength of hydrogen bond	Weak H-bonds {A,T}	Strong h-bond {G,C}

3.1.2 Lempel-Ziv (LZ) and Effort-To-Compress (ETC) complexity measures

For measuring complexity of short length segments of the nucleotide sequences, we have used Lempel-Ziv (LZ [22]) and Effort-To-Compress (ETC [37]) complexity measures. Lempel–Ziv complexity (LZ), a popular and widely used complexity measure, estimates the degree of compressibility of an input sequence. Effort-To-Compress, a more recently proposed complexity measure (by our research group), determines the number of steps required by the Non-Sequential Recursive Pair Substitution Algorithm to compress the input sequence to a constant sequence (or a sequence of zero entropy). It should be noted that both LZ and ETC are complexity measures derived from lossless data compression algorithms (hence we term them as compression-complexity measures). It has been demonstrated that both LZ and ETC outperform Shannon Entropy in characterizing complexity of noisy time series of short length arising out of stochastic (markov) and chaotic systems [37, 39, 40]. Further, ETC consistently performs better than LZ in a number of applications as shown in recently published literature [39–42]. For details of how to compute LZ and ETC on actual input sequences, we refer the readers to [22, 37, 43].

3.1.3 Distance measure and identification

We propose a very simple criteria for identification of sequences by proposing a distance measure which is computed using a compression-complexity measure (LZ or ETC). Let us say that we have genome sequences of three viruses V_1, V_2 and V_3. Firstly, we form new sequences V_1V_2 and V_2V_1 by concatenation[5]. We then compute the complexity measures $ETC(V_1)$, $ETC(V_2)$, $ETC(V_1V_2)$ and $ETC(V_2V_1)$ (similarly for LZ). In line with what has been used by Otu *et al.* [15] in Equation 4, we propose a distance measure given by the average of the relative distances between the complexity values of the two concatenated sequences V_1V_2 and V_2V_1. Mathematically, they are described as:

$$d_{LZ}(V_1, V_2) = \frac{(LZ(V_1V_2) - LZ(V_1)) + (LZ(V_2V_1) - LZ(V_2))}{2}, \tag{5}$$

$$d_{ETC}(V_1, V_2) = \frac{(ETC(V_1V_2) - ETC(V_1)) + (ETC(V_2V_1) - ETC(V_2))}{2}. \tag{6}$$

Note that the above distances will always be non-negative and symmetric[6]. In a similar fashion, we determine the distances $d_{ETC}(V_1, V_3)$ ($d_{LZ}(V_1, V_3)$) and $d_{ETC}(V_2, V_3)$ ($d_{LZ}(V_2, V_3)$). We then determine the minimum of the set $\{d(V_1, V_2), d(V_1, V_3), d(V_2, V_3)\}$[7]. We identify those viruses V_i and V_j which have the minimum distance to belong to the same group.

In our experiment, the sequences are converted to three independent sets of 0-1 sequences (by using the three methods listed in Table 2) and the LZ and ETC distances are independently calculated for all three sets. The average values of these are taken as the distance between the two sequences. In order to automatically identify the SARS viruses, the two SARS coronaviruses (Urbani and BJ01) should have the minimum distance (in complexities) as compared with the distance between the avian strain and any of the SARS coronaviruses (Urbani and BJ01). This method of identification can be easily extended if there are more than 3 sequences.

4 Results and Discussion

Figure 1 depicts the boxplots of pairwise distances (for LZ and ETC based measures) between the three viruses – Avian, BJ01 and Urbani, estimated for 100 short contiguous segments of length 30 nucleotide bases each of which were chosen independently at random locations of the entire genome sequence.

The mean pairwise distances (and standard deviations) are reported in Table 3. As it can be seen, both LZ and ETC based distance measures yield the least value for the SARS-CoV-1 virus pair.

[5] AB is the new sequence obtained by simply concatenating sequence B at the end of sequence A.
[6] $d(A, B) \geq 0$, $d(A, A) = 0$ and $d(A, B) = d(B, A)$. The triangle inequality is also likely to hold.
[7] Here $d(\cdot, \cdot)$ could be either d_{ETC} or d_{LZ}.

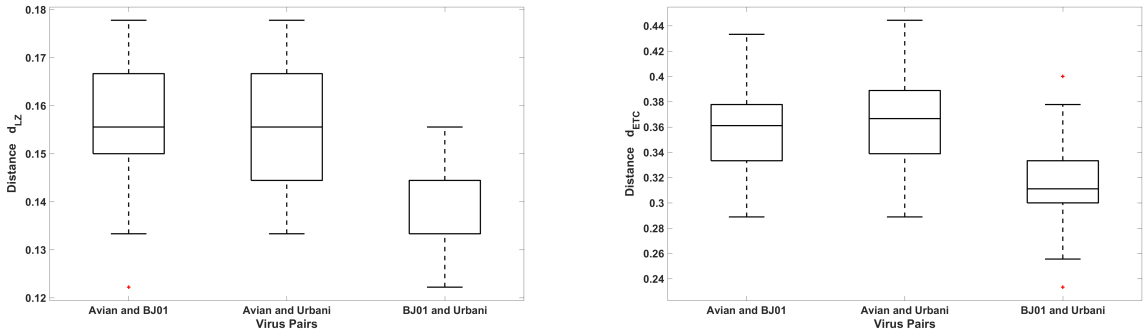

Figure 1: Boxplots of pairwise distance values for the three viruses – Avian, BJ01 and Urbani, calculated for 100 short contiguous segments of length 30 nucleotide bases chosen independently at random locations of the entire sequence. Left: d_{LZ}, Right: d_{ETC}.

Table 3: Pairwise mean distances for the three viruses – Avian, BJ01 and Urbani. We have averaged across 100 short contiguous segments of length 30 nucleotide bases each. These 100 were chosen independently at random locations of the entire genome. The mean distance between the two SARS-CoV-1 viruses BJ01 and Urbani is the least for both LZ and ETC measures.

Pair of viruses	**Distance:** d_{LZ} $(\mu \pm \sigma)$	**Distance:** d_{ETC} $(\mu \pm \sigma)$
Avian and BJ01	0.1568 ± 0.012	0.3591 ± 0.033
Avian and Urbani	0.1563 ± 0.011	0.3617 ± 0.032
BJ01 and Urbani	$\mathbf{0.1370}^{*} \pm 0.008$	$\mathbf{0.3151}^{*} \pm 0.027$

*indicates statistical significance at $p < 0.05$.

The results were statistically validated with 95% confidence interval plots as shown in Figure 2. Based on the sample data, at an overall error rate of 5%, we can conclude that both LZ and ETC are able to identify the SARS coronaviruses from the given set of viruses using only short contiguous segments consisting of 30 nucleic bases chosen independently from random locations (100 such segments) of the entire sequence.

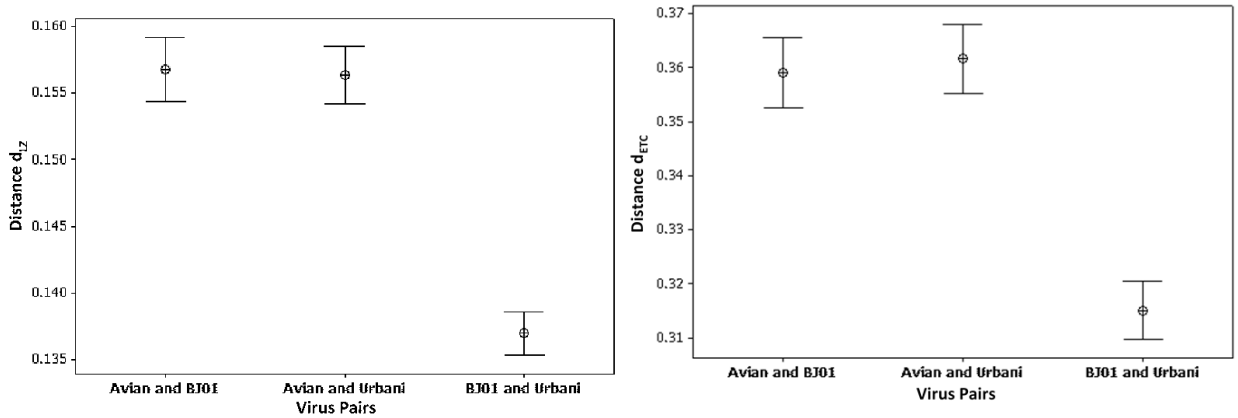

Figure 2: 95% confidence interval for mean d_{LZ} (left) and d_{ETC} (right) distance measures for the three viruses (pairwise) – Avian, BJ01 and Urbani. The distance between the two SARS-CoV-1 viruses is clearly the least for both LZ and ETC.

4.1 Distinguishing SARS-CoV-1 vs. SARS-CoV-2

Having demonstrated the efficiency of LZ and ETC based distance measures in successfully distinguishing viruses by analyzing very short segments of nucleotide sequences, we extend our work to identify SARS-CoV-2 virus from SARS-CoV-1 viruses. To this end, we use the following sequences (Table 4) for this experiment.

Table 4: Genbank accession number, name, abbreviation and length of coronaviruses used for analysis.

S.No	Accession Number	Genome	Abbreviation	Length
1	AY278741	SARS-CoV-1: Urbani	Urbani	29727
2	AY278488	SARS-CoV-1: BJ01	BJ01	29725
3	NC_004718.3	SARS-CoV-2: Reference Sequence	SARS-CoV-2	29751

Table 5: Pairwise mean distances for the three viruses – SARS-CoV-2, BJ01 and Urbani. We have averaged across 300 short contiguous segments of length 25 nucleotide bases each. These 300 segments were chosen independently at random locations of the entire genome. The mean distance between the two SARS-CoV-1 viruses BJ01 and Urbani is the least for both LZ and ETC measures. This result is statistically significant only for the ETC based measure.

Pair of viruses	Distance: d_{LZ} $(\mu \pm \sigma)$	Distance: d_{ETC} $(\mu \pm \sigma)$
SARS-CoV-2 and BJ01	0.1648 ± 0.014	0.3801 ± 0.033
SARS-CoV-2 and Urbani	0.1625 ± 0.015	0.3749 ± 0.038
BJ01 and Urbani	0.1608 ± 0.014	$\mathbf{0.3661}^{*} \pm 0.032$

*indicates statistical significance at $p < 0.05$.

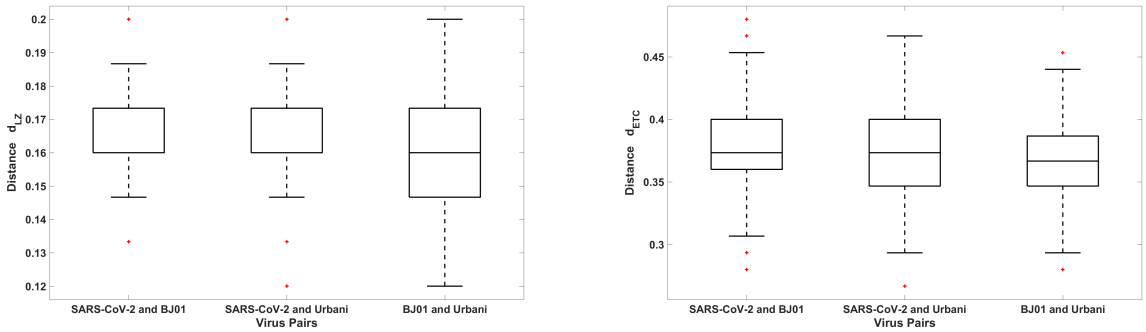

Figure 3: Boxplots of pairwise distance values for the three viruses – SARS-CoV-2, BJ01 and Urbani calculated for 300 short contiguous segments of length 25 nucleotide bases chosen independently at random locations of the entire sequence. Left: d_{LZ}, Right: d_{ETC}.

Figure 3 depicts the boxplots of pairwise distances (for LZ and ETC based measures) between the three viruses – SARS-CoV-1:BJ01, SARS-CoV-1:Urbani and SARS-CoV-2 estimated for 300 short contiguous segments of length 25 nucleotide bases each of which were chosen independently at random locations of the entire genome sequence. The mean pairwise distances (and standard deviations) are reported in Table 5. It was found that only ETC based distance measure yielded the least value (statistically significant) for the SARS-CoV-1 virus pair (BJ01 and Urbani), and not the LZ based distance measure. The statistical validation of the results is depicted using 95% confidence interval plots as shown in Figure 4. Based on the sample data, at an overall error rate of 5%, we can conclude that ETC is able to distinguish between SARS-CoV-1 and SARS-CoV-2 viruses by using only short contiguous segments consisting of 25 nucleic bases chosen independently from random locations (300 such segments) of the entire sequence. LZ based distance measure fails to do so.

To highlight the effect of segment length on distinguishing SARS-CoV-1 virsues from SARS-CoV-2 virus, we plot the pairwise distances (both LZ and ETC measures) for the three viruses for a *single* randomly chosen contiguous segment of length 5000 bases in Figure 5(left) and for another *single* randomly chosen contiguous segment of length 25 bases in Figure 5(right). It is evident that only ETC based measure is able to yield the least distance between the two SARS-CoV-1 pair of viruses for both lengths. LZ based distance measure fails to identify this pair for the short length segment and instead yields the least distance for the pair SARS-CoV-2 and Urbani which is not desirable.

Though these are still preliminary results, they are highly encouraging to further test ETC based distance measure for automatic identification/segregation of nucleotide sequences using only short segments.

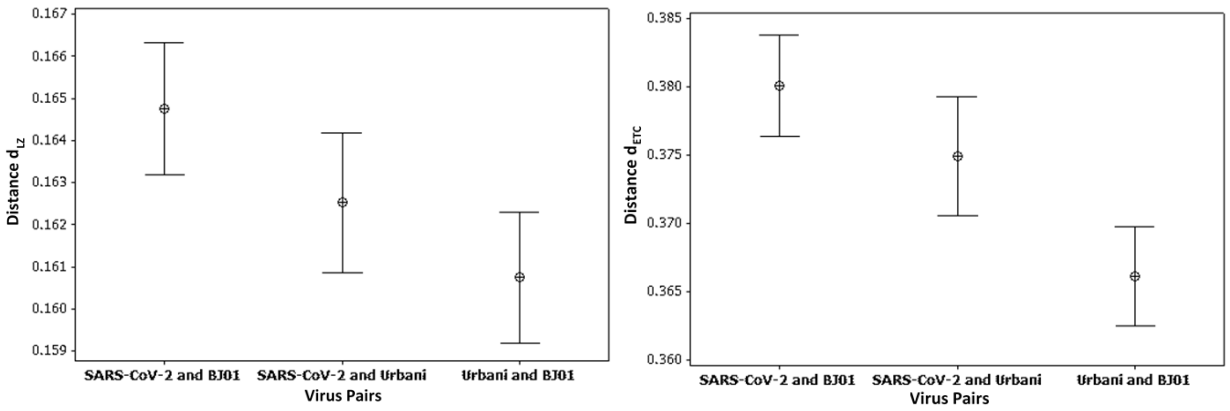

Figure 4: 95% confidence interval for mean d_{LZ} (left) and d_{ETC} (right) distance measures for the three viruses (pairwise) – SARS-CoV-2, BJ01 and Urbani. The distance between the two SARS-CoV-1 viruses is clearly the least for ETC, but not for LZ.

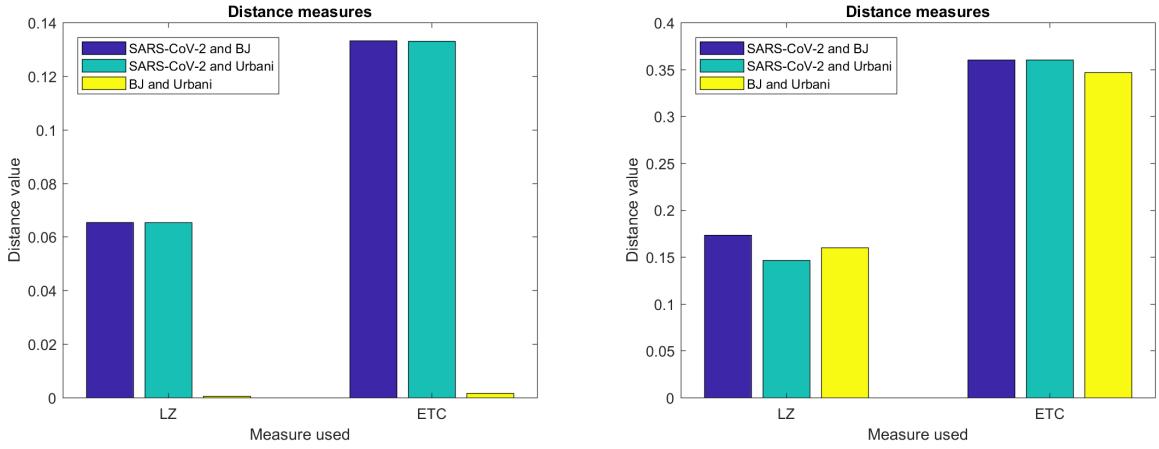

Figure 5: Effect of segment length on identification. Pairwise distances for the three viruses – SARS-CoV-2, BJ01 and Urbani computed using ETC and LZ based distance measures for a *single* randomly chosen contiguous segment of length 5000 bases (left) and 25 bases (right) from the nucleotide sequences. Only ETC correctly yields the least distance pair for the 2 SARS-CoV-1 viruses (BJ01 and Urbani) for both lengths.

5 Conclusion and Future Research Directions

Compression-complexity measures such as LZ and ETC which are based on lossless compression algorithms are good candidates for developing fast alignment-free methods for genome sequence analysis, comparison and identification. The main reason for this is their ability to characterize and analyze information in biological sequences with very short length contiguous segments. As we have demonstrated in this study, our preliminary results suggests that ETC could be very useful for identifying an unknown sequence from a large database of nucleotide sequences since we can quickly compute the measure on the candidate sequences for a small set of nucleic bases. LZ complexity requires slightly larger nucleotide sequences and that needs more computation. Other information theoretic methods in literature which employ Shannon Entropy, Mutual Information etc. would also need larger nucleotide sequences for computation and are not robust to noise. Some areas for further research are:

1. We have presented only preliminary results in this study and there is a need to rigorously test on distinguishing more sequences to further establish the reliability, robustness and universality of the proposed approach.
2. Construct a complete phylogenetic tree using the distance measure that we have proposed.
3. Compare ETC and LZ based distance measures with other methods in literature – both alignment-based and alignment-free methods.

4. Integrate ETC with existing open source packages that perform genetic sequence comparison and matching. To this end, we provide an open MATLAB® and Python implementation of ETC that can be freely downloaded and used (link provided below).

The ideas presented in this study could potentially be extended further to enable medical practitioners to rapidly and automatically identify an unknown coronavirus sample (to be either a SARS coronavirus strain or a non-SARS coronavirus) in a fast and efficient fashion using only short and/or incomplete segments of genetic sequences. Further speed up can be obtained by parallelizing the analysis on individual short segments. Potentially, the need for sequence assembly can be completely circumvented.

Software implementation of ETC

We provide open implementation of ETC (in MATLAB® and Python) for free download and use (for research and academic purposes only). Please visit: `https://sites.google.com/site/nithinnagaraj2/journal/etc`.

Acknowledgements

The authors would like to thank Gayathri R Prabhu (Indian Institute of Technology, Chennai) for helping with some of the simulations. NN would like to thank Pranay S Yadav (National Institute of Advanced Studies, Bengaluru) for the Python implementation of ETC and for useful discussions and suggestions.

References

[1] Marco A Marra, Steven JM Jones, Caroline R Astell, Robert A Holt, Angela Brooks-Wilson, Yaron SN Butterfield, Jaswinder Khattra, Jennifer K Asano, Sarah A Barber, Susanna Y Chan, et al. The genome sequence of the SARS-associated coronavirus. *Science*, 300(5624):1399–1404, 2003.

[2] Paul A Rota, M Steven Oberste, Stephan S Monroe, W Allan Nix, Ray Campagnoli, Joseph P Icenogle, Silvia Penaranda, Bettina Bankamp, Kaija Maher, Min-hsin Chen, et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*, 300(5624):1394–1399, 2003.

[3] Dylan Lebatteux, Amine M Remita, and Abdoulaye Baniré Diallo. Toward an alignment-free method for feature extraction and accurate classification of viral sequences. *Journal of Computational Biology*, 26(6):519–535, 2019.

[4] Yunxiu Zhao, Xiaolong Xue, and Xiaoli Xie. An alignment-free measure based on physicochemical properties of amino acids for protein sequence comparison. *Computational biology and chemistry*, 80:10–15, 2019.

[5] Jie Ren, Xin Bai, Yang Young Lu, Kujin Tang, Ying Wang, Gesine Reinert, and Fengzhu Sun. Alignment-free sequence analysis and applications. *Annual Review of Biomedical Data Science*, 1:93–114, 2018.

[6] Arthur Lesk. *Introduction to genomics*. Oxford University Press, 2012.

[7] Santosh Renuse, Raghothama Chaerkady, and Akhilesh Pandey. Proteogenomics. *Proteomics*, 11(4):620–630, 2011.

[8] William R Pearson. An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, 42(1):3–1, 2013.

[9] Raja Sekhar Nirujogi, Harsh Pawar, Santosh Renuse, Praveen Kumar, Sandip Chavan, Gajanan Sathe, Jyoti Sharma, Sweta Khobragade, Janhavee Pande, Bhakti Modak, et al. Moving from unsequenced to sequenced genome: reanalysis of the proteome of leishmania donovani. *Journal of proteomics*, 97:48–61, 2014.

[10] Manoj Kumar Gupta, Rajdeep Niyogi, and Mano Misra. A framework for alignment-free methods to perform similarity analysis of biological sequence. *Contemporary Computing (IC3), Sixth International Conference on*, pages 337–342, 2013.

[11] Dylan Lebatteux, Amine M Remita, and Abdoulaye Baniré Diallo. Toward an alignment-free method for feature extraction and accurate classification of viral sequences. *Journal of Computational Biology*, 26(6):519–535, 2019.

[12] Andrzej Zielezinski, Susana Vinga, Jonas Almeida, and Wojciech M Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome biology*, 18(1):186, 2017.

[13] Xuhua Xia. Distance-based phylogenetic methods. In *Bioinformatics and the Cell*, pages 343–379. Springer, 2018.

[14] Andrzej Zielezinski, Hani Z Girgis, Guillaume Bernard, Chris-Andre Leimeister, Kujin Tang, Thomas Dencker, Anna Katharina Lau, Sophie Röhling, Jae Jin Choi, Michael S Waterman, et al. Benchmarking of alignment-free sequence comparison methods. *Genome biology*, 20(1):144, 2019.

[15] Hasan H Otu and Khalid Sayood. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, 19(16):2122–2130, 2003.

[16] Stéphane Grumbach and Fariza Tahi. A new challenge for compression algorithms: genetic sequences. *Information Processing & Management*, 30(6):875–886, 1994.

[17] JS Varr, Jean-Paul Delahaye, and Eric Rivals. Transformation distances: a family of dissimilarity measures based on movements of segments. *Bioinformatics*, 15(3):194–202, 1999.

[18] Ming Li, Jonathan H Badger, Xin Chen, Sam Kwong, Paul Kearney, and Haoyong Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–154, 2001.

[19] Xin Chen, Sam Kwong, and Ming Li. A compression algorithm for dna sequences and its applications in genome comparison. In *Proceedings of the fourth annual international conference on Computational molecular biology*, page 107. ACM, 2000.

[20] LI Ming and Paul MB Vitányi. Kolmogorov complexity and its applications. *Algorithms and Complexity*, 1:187, 2014.

[21] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[22] Abraham Lempel and Jacob Ziv. On the complexity of finite sequences. *IEEE Transactions on information theory*, 22(1):75–81, 1976.

[23] Na Liu and Tian-ming Wang. A relative similarity measure for the similarity analysis of DNA sequences. *Chemical Physics Letters*, 408(4):307–311, 2005.

[24] Yi Zhang, Junkang Hao, Changjie Zhou, and Kai Chang. Normalized Lempel-Ziv complexity and its application in bio-sequence analysis. *Journal of mathematical chemistry*, 46(4):1203–1212, 2009.

[25] Bin Li, Yi-Bing Li, and Hong-Bo He. LZ complexity distance of DNA sequences and its application in phylogenetic tree reconstruction. *Genomics Proteomics & Bioinformatics*, 3(4):206–212, 2005.

[26] Liwei Liu, Dongbo Li, and Fenglan Bai. A relative Lempel-Ziv complexity: Application to comparing biological sequences. *Chemical Physics Letters*, 530:107–112, 2012.

[27] Chenglong Yu, Rong Lucy He, and Stephen S-T Yau. Viral genome phylogeny based on lempel–ziv complexity and hausdorff distance. *Journal of theoretical biology*, 348:12–20, 2014.

[28] Yong-Joon Song and Dong-Ho Cho. Classification of various genomic sequences based on distribution of repeated k-word. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3894–3897. IEEE, 2017.

[29] Ricardo E Monge and Juan L Crespo. Comparison of complexity measures for DNA sequence analysis. *Bio-inspired Intelligence (IWOBI), International Work Conference on*, pages 71–75, 2014.

[30] Khalid Sayood, Hasan H Otu, and Steven H Hinrichs. System and method for sequence distance measure for phylogenetic tree construction, May 13 2014. US Patent 8,725,419.

[31] Tiee-Jian Wu, Ya-Ching Hsieh, and Lung-An Li. Statistical measures of dna sequence dissimilarity under markov chain models of base composition. *Biometrics*, 57(2):441–448, 2001.

[32] B Edwin Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences*, 83(14):5155–5159, 1986.

[33] Zurab Bzhalava, Emilie Hultin, and Joakim Dillner. Extension of the viral ecology in humans using viral profile hidden markov models. *PloS one*, 13(1), 2018.

[34] Tuan D Pham and Johannes Zuegg. A probabilistic measure for alignment-free sequence comparison. *Bioinformatics*, 20(18):3455–3461, 2004.

[35] Chenglong Yu, Mo Deng, and Stephen S-T Yau. DNA sequence comparison by a novel probabilistic method. *Information Sciences*, 181(8):1484–1492, 2011.

[36] Masooda Omari, Tyler W Barrus, Mark Sanders, and Daniel Negron. Rapid genomic sequence classification using probabilistic data structures, November 15 2018. US Patent App. 15/977,667.

[37] Nithin Nagaraj, Karthi Balasubramanian, and Sutirth Dey. A new complexity measure for time series analysis and classification. *The European Physical Journal Special Topics*, 222(3-4):847–860, 2013.

[38] Ping-an He and Jun Wang. Characteristic sequences for dna primary sequence. *Journal of chemical information and computer sciences*, 42(5):1080–1085, 2002.

[39] Nithin Nagaraj and Karthi Balasubramanian. Dynamical complexity of short and noisy time series. *The European Physical Journal Special Topics*, 226(10):2191–2204, 2017.

[40] Nithin Nagaraj and Karthi Balasubramanian. Three perspectives on complexity: entropy, compression, subsymmetry. *The European Physical Journal Special Topics*, 226(15):3251–3272, 2017.

[41] Marjola Thanaj, Andrew J Chipperfield, and Geraldine F Clough. Multiscale analysis of microvascular blood flow and oxygenation. In *World Congress on Medical Physics and Biomedical Engineering 2018*, pages 195–200. Springer, 2019.

[42] Nithin Nagaraj and Karthi Balasubramanian. Measuring complexity of chaotic systems with cybernetics applications. In *Handbook of Research on Applied Cybernetics and Systems Science*, pages 301–334. IGI Global, 2017.

[43] Virmani Mohit and Nithin Nagaraj. A novel perturbation based compression complexity measure for networks. *Heliyon 5-e01181*, 5(2), 2019.