Genome-wide Analysis to Identify Palindromes, Mirror and Inverted Repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1

N.GHOSH¹, I. SAHA² and D. PLEWCZYSNKI³,⁴.
¹Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha ‘O’ Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India (e-mail: ghosh.nimisha@gmail.com)
²Department of Computer Science and Engineering, National Institute of Teachers’ Training and Research, Kolkata, West Bengal, India (e-mail: indrajit@nitttrkol.ac.in)
³Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland (e-mail: d.plewczynski@mini.pw.edu.pl)
⁴Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland
Corresponding author: I. Saha (e-mail: indrajit@nitttrkol.ac.in) and D. Plewczynski (e-mail: d.plewczynski@mini.pw.edu.pl)

This work has been partially supported by CRG short term research grant on COVID-19 (CVD/2020/000991) from Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India. However, they do not provide any publication fees. This work is co-funded by Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme. This work has also been co-supported by Polish National Science Centre (2019/35/O/ST6/02484 and 2020/37/B/NZ2/03757).

ABSTRACT Research pertaining to SARS-CoV-2 is in full swing to understand the origin and evolution of this deadly virus that can lead to its rapid detection. To achieve this, atypical genomic sequences which may be unique to SARS-CoV-2 or Coronaviridae family in general may be investigated. Such sequences in virus genomes may be responsible for target prediction, replication, defence mechanisms and viral packaging. This fact has motivated us to explore the different types of repeats such as palindromes, mirror repeats and inverted repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1. For this purpose, the respective reference sequence of SARS-CoV-2, MERS-CoV and SARS-CoV-1 is divided into descriptors of sequences of length \(k\) using \(k\)-mer technique. Thereafter, these descriptors are represented as a collection of tokens which are subsequently used for the identification of palindrome, mirror repeat and inverted repeat in the respective reference sequence. The highest number of palindromes, mirror repeats and inverted repeats are identified for descriptor length 10. As a result, for palindromes such values are 38, 42 and 33 and for mirror repeats they are 52, 38 and 33 for SARS-CoV-2, MERS-CoV and SARS-CoV-1 respectively. For inverted repeats, with a descriptor length 10 and intervening length 5, the values are 59, 56 and 70 respectively. Moreover, the identified repeats are then searched for in 108246, 291 and 340 SARS-CoV-2, MERS-CoV and SARS-CoV-1 virus sequences respectively to find the population coverage of such repeats. It surpasses 99% in most cases and even 100% for some. Furthermore, GC contents which mostly lie between 20%-50% are evaluated for these repeats as well in order to understand their binding efficacy.

INDEX TERMS COVID-19, Inverted Repeats, Mirror Repeats, Palindromes, SARS-CoV-2

I. INTRODUCTION COVID-19, Inverted Repeats, Mirror Repeats, Palindromes, SARS-CoV-2

The outbreak of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), the virus causing COVID-19, has caused more than five million deaths in different parts of the world leading to researches around the globe to control this virus. The most common symptoms of COVID-19 include fever, dry cough, dyspnea, headache and pneumonia and the less common symptoms are gastrointestinal symptoms, myalgias, rhinorrhea and chest pain [1]; the fatality rate is high in patients suffering from cancer, diabetes or cardiovascular disorders [2]. SARS-CoV-2 belongs to the family of Coronaviridae which also houses MERS-CoV and SARS-CoV-1 viruses [1], [3]. Research pertaining to SARS-CoV-2 is in full swing to understand the origin and evolution of this
deadly virus. This may lead to the immune-related studies of viruses [4]. To achieve this, atypical genomic sequences which may be unique to SARS-CoV-2 or Coronaviridae family in general may be investigated [5]–[8]. Studies [9]–[11] have suggested that such sequences in virus genomes may be responsible for target prediction, replication, defence mechanisms and viral packaging. In this regard, Goswami et al. [6] have identified inverted repeats in hotspot mutations using Palindrome analyzer webserver. Siddique et al. [12] have also identified simple sequence repeats using web tools like Sequence Repeat Identification Tool and IMEx-web: Imperfect Microsatellite Extraction Webserver. From the biological perspective as well, repeats are correlated with several phenomena. They are known to form functional association between genes in genetic networks and co-expression of genes as well as an entire gene or operon [13]. Moreover, inverted repeats have many important biological functionalities that play significant role in genome instability and may lead to mutation and disease.

The aforementioned studies have motivated us to explore the different types of repeats such as palindromes, mirror repeats and inverted repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1. A palindromic sequence is a symmetrical sequence so that when read from the reverse direction, it is the exact complement of itself. For example, TGCA is a palindrome of length 4. It is to be noted that a palindrome is always even in length. On the other hand, a mirror repeat has inverted sequence occurring within individual strands of DNA. For example, TTAGGATT is a mirror repeat. Like palindromes, mirror repeats are also even in length. An inverted repeat sequence is followed downstream by its reverse complement and has intervening length of sequence that can vary from zero to any number. If the intervening length is zero, it is a palindromic sequence. For example, CGTAAxxxxxTTACG is an inverted repeat, where xxxxxx are the intervening sequence of length 6. If the intervening length is zero, CGTAAATTACG is a palindrome. The significance of each of the repeats is many folds. As palindromic sequence reads same in both directions, it is important for the replication, repair mechanisms and transcription. These repeated sequence patterns are thus widely used for their fundamental importance in understanding the genome function and organization [13].

In this work, reference sequences of SARS-CoV-2, MERS-CoV and SARS-CoV-1 are divided into descriptors of sequences of length k using the popular k-mer technique. Thereafter, these descriptors are represented as a collection of tokens which are subsequently used for the identification of palindromes, mirror and inverted repeats in the respective reference sequence. In this regard, such repeats of length 10 to 20 are identified in the reference sequence of SARS-CoV-2, MERS-CoV and SARS-CoV-1. Subsequently, the identified repeats are then searched for in 108246, 291 and 340 SARS-CoV-2, MERS-CoV and SARS-CoV-1 virus sequences respectively to find their population coverage. Moreover, the GC contents are also evaluated for these repeats in order to understand their binding efficacy. Furthermore, repeats are identified in the mutated variants of the virus like Alpha, Beta, Delta and Gamma as well. It is worth mentioning that we have used our own method in this work to identify the different repeats like palindrome, mirror repeat and inverted repeat as contrary to the ones mentioned in the literature like [6] and [12] who have used online tools for the identification of the repeats. Also, please note that, although in [14] the author has identified palindromes in SARS-CoV-2 genomes, it is not biologically appropriate and this has also motivated us to correctly identify different types of repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1. Also, palindromic sequences can be identified by using tools like http://www.biophp.org/minitools/find_palindromes/demo.php, https://www.bioinformatics.nl/cgi-bin/emboss/palindrome, https://www.novoprolabs.com/tools/dna-palindrome and that reported in [13] and [15]. However, from the perspective of SARS-CoV-2, our work is first of a kind which provides a comprehensive idea of palindromes, mirror and inverted repeats. Moreover, the code used to identify such repeats is free to use as well for the readers which is not the case for many of the works in the literature.

II. MATERIALS AND METHODS

In this section, the various terms used in this work, the collection of the datasets for SARS-CoV-2, MERS-CoV and SARS-CoV-1 genomes and the proposed pipeline are discussed.

A. BACKGROUND

Before delving into the methodology, the several terms used in this work are briefly discussed. The motivation of this work is to find palindromic sequences, mirror and inverted repeats for SARS-CoV-2, MERS-CoV and SARS-CoV-1. For the benefit of the readers, the description of palindromes, mirror and inverted repeats are already introduced in the Introduction section. Furthermore, descriptors are small segments or subsequences of length k in which the reference sequence of each of the virus is divided to identify the different repeats as mentioned in this work.

B. DATA ACQUISITION

The pipeline of the work is given in Figure 1(a). To find the palindromes, mirror repeats and inverted repeats, initially three reference genomic sequences of SARS-CoV-2 (NC_045512.2) [2] MERS-CoV (NC_019843.3) [3] and SARS-CoV-1 are considered (NC_004718.3) which are collected from NCBI (National Center for Biotechnology Information). Once the aforementioned repeats are identified for each virus on their respective reference sequence, their population coverage is verified on 108246, 291 and 340 virus sequences SARS-CoV-2, MERS-CoV and SARS-CoV-1. The virus

1https://www.ncbi.nlm.nih.gov/nuccore/1798174254
2https://www.ncbi.nlm.nih.gov/nuccore/667489388
3https://www.ncbi.nlm.nih.gov/nuccore/30271926
sequences for SARS-CoV-2 are collected from Global Initiative on Sharing All Influenza Data (GISAID)\(^1\) in fasta format while the sequences for MERS-CoV and SARS-CoV-1 are collected from NCBI\(^2\).

C. PIPELINE OF THE WORK

This work is executed according to the pipeline as given in Figure 1(a) and Figure 1(b) provides a demonstration of the workflow.

Algorithm 1 presents the workflow as provided in Figure 1(b) for the identification of a palindromic sequence, mirror repeat and inverse repeat in details. To identify each such repeat, the reference sequence \(\chi\) for each of the virus is initially divided into subsequences or descriptors of length \(k\) using the popular \(k\)-mer technique and stored in \(C\). Thereafter, these descriptors are represented as a collection of words or tokens using the function \(\text{tokenizedDocument}\) and stored in \(\mathcal{P}\). Subsequently, these tokens are checked if they qualify to be either palindromic sequence, mirror repeat or inverted repeat. To perform such operation for palindrome, the reverse complement of the tokens stored in \(\mathcal{P}\) are calculated. For those tokens, if the original token and their reverse complement are same, then they are palindromes. To find the mirror repeat, the tokens are flipped in the horizontal direction. For those tokens whose original and flipped version are same, they are considered to be the mirror repeats.

Finally, for the inverse repeat, initially, the starting positions (\(\text{start\_pos}\_F\)) of the tokens in the reference sequence are calculated. From these starting positions, the starting positions (\(\text{start\_pos}\_rev\)) of the probable reverse complements are calculated by considering \(k\) and \(\eta\) which is the intervening length for inverted repeat. Based on \(\text{start\_pos}\_rev\) and \(\text{end\_pos}\_rev\), the intermediate sequences between these two positions are extracted from the reference sequence \(\chi\). If these tokens and the reverse complements of the original tokens match, then the original tokens are considered to be inverted repeats. Once the repeats are identified, their population coverage is calculated for all the virus sequences. Furthermore, the GC content of the identified repeats are also reported.

The time complexity of the proposed method for finding the palindrome, mirror repeats and inverted repeats as given in Algorithm 1 can be calculated as: Let the length of the reference sequence be \(N\). Thus, the time complexity of finding descriptors of sequences of repeats using \(\text{kmercount}\) is \(O(N)\) (Line number 3). The time complexity of navigating through the sequence of descriptors as returned by \(\text{kmercount}\) is \(O(\alpha)\), where \(\alpha = N - k + 1\) is the number of descriptors. So, the time complexity for representing the descriptors as tokens is \(O(\alpha k)\) (Line number 4) \((k\) is the length of a particular descriptor out of \(\alpha\) descriptors). The time complexity of navigating through the tokens is \(O(\alpha)\) as well. Line 7 has a time complexity of \(O(k)\) and so do lines 8 and 11 as well. Thus, the total complexity of the proposed algorithm is \(O(N + \alpha k + \alpha k) \approx O(N + \alpha k)\). On the other hand, the required memory is as follows: The memory needed for \(\text{kmercount}\) is \(N + \alpha \times k\) making the space complexity as \(O(N) + O(\alpha) \ast k\) and that for \(\text{tokenizedDocument}\) is \(O(2N)\). Finally, for \(\text{reversecomplement}\), \(\text{flip}\) and \(\text{extractBetween}\) the space required is \(O(3k)\). Thus, the total space complexity is \((O(N) + O((\alpha) \ast k)) + O(2N) + O(3k))\).

III. RESULTS

The experiments in this work are conducted using MATLAB R2021a and the results following Figure 1(a) are given next. Figure 1(b) provides a glimpse of each of the repeats as identified in this work. As can be seen from the figure, ACCAGT-TAACTGTG is a palindrome because when seen from the reverse direction, it gives TGGTCATTGACCA which is an exact complement of itself. AATGACTTCAGTAA is a mirror repeat as the last 7 nucleotides TCAGTAA is a mirror image of the first 7 that is, AATGACT. Finally, AA-CACCTTCxxxxxGAAGTGTGTTT is an example of an inverted repeat as GAAAGTGTGTTT is an exact complement of AACACTTC with an intervening length of 5 (marked by x). The results for the total number of palindromes and mirror repeats in the reference genomic sequence of each of SARS-CoV-2, MERS-CoV and SARS-CoV-1 are shown in Table 1 where the lengths of the descriptors are varied from 10 to 20, considering all even lengths. For SARS-CoV-2 with a length of 10, the number of palindromes are 38 while for MERS-CoV and SARS-CoV-1, the number of such sequences are 42 and 33 respectively while the number of mirror repeat for the

Algorithm 1: Pseudo-code for identification of palindromic sequence, mirror repeat and inverse repeat

| Line | Description |
|------|-------------|
| 1    | Initialize NULL Sets, \(S_{PAL}, S_{MR}\), and \(S_{IV}\) |
| 2    | \(C \leftarrow \text{kmercount}(\chi, k)\) |
| 3    | \(\mathcal{P} \leftarrow \text{tokenizedDocument}(C)\) |
| 4    | for \(i = 1\) to length(\(\mathcal{P}\)) do |
| 5    | \(F \leftarrow \mathcal{P}(i, 1)\) |
| 6    | \(R_{PAL} = \text{reversecomplement}(F)\) |
| 7    | \(R_{MR} = \text{flip}(F)\) |
| 8    | \(\text{start\_pos}_F = \text{start\_pos}_F + k + \eta\) |
| 9    | \(\text{end\_pos}_rev = \text{start\_pos}_rev + (k-1)\) |
| 10   | \(\text{S}_{IV} = \text{extractBetween}(\chi, \text{start\_pos}_rev, \text{end\_pos}_rev)\) |
| 11   | if \(F == R_{PAL}\) then |
| 12   | \(\mathcal{P}_{S_{PAL}} = \mathcal{P}_{S_{PAL}} \cup F\) |
| 13   | else if \(F == R_{MR}\) then |
| 14   | \(\mathcal{P}_{S_{MR}} = \mathcal{P}_{S_{MR}} \cup F\) |
| 15   | else if \(R_{PAL} == S_{IV}\) then |
| 16   | \(\mathcal{P}_{S_{IV}} = \mathcal{P}_{S_{IV}} \cup F\) |
| 17   | continue |
| 18   | end |

---

\(^1\)https://gisaid.org/CoV2020
\(^2\)https://www.ncbi.nlm.nih.gov/genome/viruses
Input the reference genomic sequences of SARS-CoV-2, MERS-CoV and SARS-CoV-1 from NCBI

Verify their population coverages and find their GC contents for 108,246 SARS-CoV-2, 291 MERS-CoV and 340 SARS-CoV-1 sequences

Palindromic Sequence
ACCAGTTAACTGGT

Complementary Sequence
TGGTCAAATTGACCA

Mirror Repeat
AATGACTTCAGTAA

Inverted Repeat
ACACTTCxxxxxGAAGTGT

FIGURE 1: (a) Pipeline of the Work, (b) Demonstration of workflow (c) Number of Palindromic Sequences, Mirror Repeats and Inverted Repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1 viruses and (d) Venn Diagrams to show the common Palindromic Sequences, Mirror Repeats and Inverted Repeats among SARS-CoV-2, MERS-CoV and SARS-CoV-1 viruses for \( k = 10 \)
same length are 52, 38 and 33 respectively for SARS-CoV-2, MERS-CoV and SARS-CoV-1. For MERS-CoV, there are no palindromes beyond the length of 18 and no mirror repeats beyond the length of 14. On the other hand, for SARS-CoV-1, there are no mirror repeats beyond the length of 16. For all the three viruses, there are neither palindromes nor mirror repeats over the length of 20. The results for inverted repeats for all the three viruses are given in Table 2. The intervening length for the inverted repeats has been varied from 5 to 8. For SARS-CoV-2, inverted repeats are present only for lengths 18 and 20, with intervening lengths of 5 and 7, while for MERS-CoV and SARS-CoV-1, such sequences are present for length 18 and intervening length 5. The results for palindromes along with the corresponding coding gene, population coverage and GC content are given in Table 3 with descriptor length varying from 14 to 20. For length = 20, the palindrome for SARS-CoV-2 is ACACTGGTAATTACCAGTGT and the corresponding genomic location in the reference sequence is 5745, the coding gene being NSP3. The corresponding population coverage and GC content are 99.82% and 40% respectively. For MERS-CoV, with a descriptor length of 18, the palindromic sequence is CTATAGAGATCTCTATAG at the genomic location 16171 and it belongs to the coding gene RNA-dependent RNA polymerase with a population coverage and GC content of 28.52% and 33.33% respectively. On the other hand, for SARS-CoV-1, for a length of 18, the palindrome is CTTTAACAAGCTTGTTAAG at the genomic location 25963 and belongs to the coding gene ORF3a. The corresponding population coverage and GC content are 91.17% and 30% respectively.

### Table 1: Number of Palindromes and Mirror Repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1 Viruses

| Length of Descriptors | Number of Palindromes | Number of Mirror Repeats |
|------------------------|-----------------------|--------------------------|
| SARS-CoV-2             | MERS-CoV               | SARS-CoV-1               |
| 14                     | 15                    | 12                       |
| 12                     | 11                    | 12                       |
| 14                     | 5                     | 4                        |
| 16                     | 1                     | 1                        |
| 20                     | 1                     | 0                        |

### Table 2: Number of Inverted Repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1 Viruses

| Virus         | Length of Descriptors (k) | Intervening Length (η) |      |      |
|---------------|---------------------------|------------------------|------|------|
| SARS-CoV-2    | 10                        | 5, 6, 7, 8             | 59   | 59   |
| MERS-CoV      | 10                        | 56, 56                 | 63   | 51   |
| SARS-CoV-1    | 10                        | 56, 56                 | 63   | 51   |

### Table 3: Palindromes in SARS-CoV-2, MERS-CoV and SARS-CoV-1 Viruses along with the corresponding coding gene, population coverage and GC content

### IV. DISCUSSION

The findings in this work are solely based on the proposed technique. In this regard, apart from the identification of the repeats, their population coverage and GC content are also elaborated. Table 3 reports the population coverage and the GC content for each of the identified palindromes. Population coverage shows the percentage of coverage of the identified palindromes, mirror repeats and inverted repeats in 108246, 291 and 340 virus sequences of SARS-CoV-2, MERS-CoV and SARS-CoV-1. For example, ACACCTGGTAATTACCAGTGT is present in 108052 out of 108246 sequences, resulting in a population coverage of 99.82%. On the other hand, GC content provides the number of guanine and cytosine bases in the identified palindromes. For example, ACACCTGGTAATTACCAGTGT has a GC content of 28.52% and 33.33% respectively. For
length = 20, the mirror repeat for SARS-CoV-2 is CTCAAT-GACTTCAGTAACT and the corresponding genomic location in the reference sequence is 9977. The coding gene for the sequence is NSP4 with a population coverage of 99.56% and GC content of 40%. For MERS-CoV, for a mirror repeat length of 14, one of the the sequences is CAGATGGTGT-GAC at the genomic location 12574 and it belongs to the coding gene NSP8 while for SARS-CoV-1, for a length of 16, the mirror repeat is CTACTGACCAGTCATC at the genomic location 7643 and belongs to the coding gene NSP3.

For inverted repeats with a length of 20 and intervening length of 7, as reported in Table 3, the SARS-CoV-2 sequence is ATAAAGACTxxxxxAAGTTTCTTTAT where ‘xxxxx’ denotes the intervening sequence of length 7. In this case the intervening sequence is TTAAGTC. The sequence ATAAAGACTTTAATGCACTTTTAT (by replacing TTAAGTC in place of x’s in ATAAAGAC-TxxxxxAAGTTTCTTTAT) has the starting coordinate 15775 and belongs to the coding gene RNA-dependent RNA polymerase with a population coverage of 99.88% and GC content of 22.22%. According to [10, 11], targeting GC-rich genes are difficult and thus it can be said that repeats with moderate GC-content are good candidates for being target sites of viruses. Thus, the identified repeats in this work can be considered to be such good targets. Examples for each of the repeats are provided in Figure 1(b) as well where ACCAT-TAACGTG is a palindrome, AATGACCTCAAGTA is a mirror repeat and ACACCTCxxxxxGAGTGT is an inverted repeat where the intervening sequence of length 5 is CCACA (marked in orange in the figure). Figure 1(c) provides the pictorial representations of Tables 1 and 2 where ‘5’, ‘6’, ‘7’ and ‘8’ are the intervening lengths (η) of the inverted repeats. Figure 1(d) shows the common repeats among SARS-CoV-2, MERS-CoV and SARS-CoV-1 for descriptor length (k) 10. It can be seen from the figure that although some of the repeats are similar, mostly they are different. The similarity may be attributed to the fact that they all belong to the same family of Coronaviridae but due to evolutionary pressure their structures have changed. All the results for the identified palindromes, mirror repeats and inverted repeats of SARS-CoV-2, MERS-CoV and SARS-CoV-1 along with the coding gene, population coverage and GC content are given in the supplementary as excel files.

Furthermore, major mutations of the original reference sequence of the virus have led to important variants like Alpha, Beta, Delta and Gamma. We have also identified the palindromes, mirror repeats and inverted repeats in such variants. As can be seen from Figure 2, most of such repeats are common among all the variants. In Figure 2(a), there are 38 common palindromes of length 10 among the reference sequence and all the variants where ACCATATGGT is the only extra palindrome present in Alpha variant. In Figure 2(b), number of common mirror repeats of length 10 is 49 among all while 2 (CAGGTGTGAGC and TTTCGGCTTT) are common among Reference, Alpha Beta and Gamma and 3 (AACATGTTGTA, ATATAATATA and AACAAACCAA) are common among Reference, Alpha and Delta variants like TTACCAGTTG, GAC at the genomic location 12574 and it belongs to the coding gene NSP4 with a population coverage of 99.56% and GC content of 40%. For MERS-CoV, for a length of 14, there are 59 such sequences common among all while only 1 (CACAACACATTTGGT) is unique in Alpha variant which is depicted in Figure 2(c). For intervening length of 6 and 7, there are 69 and 61 common inverted repeats among all as shown in Figs. 2(d) and (e) respectively while there is 1 unique repeat (ACACCATTATTGTG) is unique in Alpha variant and 1 (ACACCTTGAATTTGTG) common among Reference, Alpha, Beta and Gamma. Finally, as given in Figure 2(f) for intervening length of 8, there are 59 common inverted repeats among the Reference and all the other mutated variants. The rest of the results are provided in the supplementary.

### V. LIMITATIONS OF THE STUDY

Extensive experiments have been performed in this study to find palindromes, mirror and inverted repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1. However, the study has certain limitations. First, there are about 7 species of coronavirus like 229E, NL63, OC43, HKU1, MERS-CoV, SARS-CoV-1, SARS-CoV-2 that transmit from animals to humans. But in this study, we have considered SARS-CoV-2, MERS-CoV and SARS-CoV-1. The study can be extended to other species as well which we intend to do in our future work. Second, in vivo-based verification can further add significance to the study. We intend to collaborate with biologists in the future to perform the same.

### VI. CONCLUSION

SARS-CoV-2 and its different variants have been wreaking havoc in everybody’s life for the past two years. Thus, their rapid detection is one of the most potent ways to keep the virus somewhat in control. In this regard, atypical genomic sequences unique to SARS-CoV-2 or Coronavirus family in general can be explored to fulfill such purpose. Keeping this in context, palindromes, mirror and inverted repeats are identified in SARS-CoV-2, MERS-CoV and SARS-CoV-1 viruses along with their population coverage and GC content. GC contents are evaluated for the identified repeats in order

| Value | Length (k) | Description (η) | Mirror Repeat Sequence | Start Coordinate of Mirror Repeat | Coding Gene | Population Coverage (%) | GC Content (%) |
|-------|------------|-----------------|------------------------|-------------------------------|-------------|--------------------------|----------------|

| SARS-CoV-2 | 14 | ATACATGTTGTA | NSP4 | 20.66 | 22.57 |
|------------|----|---------------|------|-------|------|
| 15 | ATACATGTTGTA | NSP4 | 20.66 | 22.57 |
| 40 | CTCAATGACTTC | NSP8 | 99.86 | 99.86 |

| MERS-CoV | 14 | CTCAATGACTTC | NSP4 | 99.86 | 99.86 |
|----------|----|---------------|------|-------|------|
| 20 | CTCAATGACTTC | NSP4 | 99.86 | 99.86 |

| SARS-CoV-1 | 14 | CTCAATGACTTC | NSP4 | 99.86 | 99.86 |
|------------|----|---------------|------|-------|------|
| 15 | CTCAATGACTTC | NSP4 | 99.86 | 99.86 |
to understand the binding efficacy with the corresponding complements while targeting such repeats for detecting the virus. Hence, using the identified repeats further research can be conducted in this direction as well. Moreover, such repeats are identified in major mutations of SARS-CoV-2 like Alpha, Beta, Delta and Gamma and it is seen that most of the repeats being conserved in nature are common across the original sequence and the different variants. As a future work, parameterized matching [18] can be considered to identify repeats.

### AVAILABLE DATA AND MATERIALS

All the SARS-CoV-2, MERS-CoV and SARS-CoV-1 virus genomes with their corresponding reference sequences and the final results of this work are available at “http://www.nitttrkol.ac.in/indrajit/projects/COVID-Repeats/”.

### ACKNOWLEDGMENT

We thank all those who have contributed sequences to GI-SAID and NCBI databases.

### REFERENCES

[1] P. Zhou, X. L. Yang, X. G. Wang, and et al., “A pneumonia outbreak associated with a new coronavirus of probable bat origin,” *Nature*, vol. 579, pp. 270–273, 2020. doi: 10.1038/s41586-020-2012-7

[2] B. Chakrabarty, D. Das, G. Bulusu, and et al., “Network-based analysis of fatal comorbidities of COVID-19 and potential therapeutics,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 4, pp. 1271–1280, 2021. doi: 10.1109/TCBB.2021.3075299

[3] Z. Chen, S. S. Boon, M. H. Wang, and et al., “Genomic and evolutionary comparison between SARS-CoV-2 and other human coronaviruses,” *Journal of Virological Methods*, vol. 289, p. 114032, 2021. doi: 10.1016/j.jvirohet.2020.114032

[4] P. A. Rota, M. S. Oberste, S. S. Monroe, and et al., “Characterization of a novel coronavirus associated with severe acute respiratory syndrome,” *Science*, vol. 300, no. 5624, pp. 1394–1399, 2003. doi: 10.1126/science.1085952

[5] M. Bartas, V. Brázdí, N. Bohdálová, and et al., “In-depth bioinformatic analyses of nidovirales including human SARS-CoV-2, SARS-CoV,

---

**TABLE 5: Inverted Repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1 viruses along with the corresponding coding gene, population coverage and GC content**

| Virus          | Length of Description (A) | Intervening Length (nt) | Intervening Repeats Sequence | Intervening Sequence | Start Coordinate of Intervened Repeats | Coding Gene  | Population Coverage (%) | GC Content (%) |
|---------------|--------------------------|-------------------------|-----------------------------|---------------------|---------------------------------------|--------------|-------------------------|----------------|
MERS-CoV viruses suggest important roles of non-canonical nucleic acid structures in their lifecycles; *Frontiers in Microbiology*, vol. 11, p. 1583, 2020. doi: 10.3389/fmicb.2020.01583

[6] P. Goswami, M. Bartas, M. Lexa, and et al., “SARS-CoV-2 hot-spot mutations are significantly enriched within inverted repeats and CpG island loci,” *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1338–1345, 12 2020. doi: 10.1093/bib/bbaa385

[7] D. Ji, M. Juhas, C. M. Tsang, and et al., “Discovery of G-quadruplex-forming sequences in SARS-CoV-2,” *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1150–1160, 2020. doi: 10.1093/bib/bbaa14

[8] H. Cui and Z. L., “G-Quadruplexes are present in human coronaviruses including SARS-CoV-2,” *Frontiers in Microbiology*, vol. 11, 2020. doi: 10.3389/fmicb.2020.567317

[9] D. Cain, O. Erwine, A. Grigg, and et al., “Palindromic sequence plays a critical role in human foamy virus dimerization,” *Journal of Virology*, vol. 83, no. 8, pp. 3731–3739, 01 2007. doi: 10.1128/JVI.03731-01

[10] A. M. D. Dirac, H. Huthoff, J. Kjems, and et al., “Requirements for RNA heterodimerization of the human immunodeficiency virus type 1 (HIV-1) and HIV-2 genomes,” *Journal of General Virology*, vol. 83, no. 10, pp. 2533–2542, 2002. doi: 10.1099/0022-1317-83-10-2533

[11] D. Chew, K. P. Choi, H. Heidner, and et al., “Palindromes in SARS and other coronaviruses,” *INFORMS Journal on Computing*, vol. 16, no. 4, pp. 331–340, 11 2004. doi: 10.1287/ijoc.1040.0087

[12] R. Siddique and A. Ghosh, “Genome-wide in silico identification and characterization of complex sequence repeats in diverse completed SARS-COV-2 genomes,” *Gene Reports*, vol. 23, p. 101020, 2021. doi: 10.1016/j.genrep.2021.101020

[13] M. Gurusaran, D. Ravella, and K. Sekar, “RepEx: Repeat extractor for biological sequences,” *Genomics*, vol. 102, no. 4, pp. 403–408, 2013. doi: 10.1016/j.ygeno.2013.03.005

[14] W. R. Gallaher, “A palindromic RNA sequence as a common breakpoint with breakpoint contribution to copy-choice recombination in SARS-COV-2,” *Archives of Virology*, vol. 165, no. 10, pp. 2341–2348, 2020. doi: 10.1007/s00705-020-04750-z

[15] M. Berselli, E. Lavez, and S. Toppo, “NeSSie: a tool for the identification of approximate DNA sequence symmetries,” *Bioinformatics*, vol. 34, no. 14, pp. 2503–2505, 2018. doi: 10.1093/bioinformatics/bty142

[16] M. Haeussler, K. Schönig, H. Eckert, and et al., “Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR,” *Genome Biology*, vol. 17, no. 148, 2016. doi: 10.1186/s13059-016-1012-2

[17] A. Reynolds, D. Leake, Q. Boese, and et al., “Rational siRNA design for RNA interference,” *Nature Biotechnology*, vol. 22, p. 326–330, 2004. doi: 10.1038/nbt836

[18] F. Cauterucci, G. Terracina, and D. Ursino, “Generalizing identity-based string comparison metrics: Framework and techniques,” *Knowledge-Based Systems*, vol. 187, p. 104820, 2020. doi: 10.1016/j.knosys.2019.06.028

**FIGURE 2:** Venn Diagram to show the common (a) Palindromic Sequences, (b) Mirror Repeats and Inverted Repeats for $k = 10$ and (c) $\eta = 5$, (d) $\eta = 6$, (e) $\eta = 7$ and (f) $\eta = 8$ among the Reference (original virus) and the different variants of SARS-CoV-2
NIMISHA GHOSH has completed her Masters and PhD from IIEST, Shibpur, India before starting as an Assistant Professor in Siksha O Anusandhan (Deemed to be) University on Bhubaneswar, India. She has more than 30 publications in reputed journals and conferences. Her research interests include Computational Biology, Machine Learning and Wireless Sensor Networks.

INDRAJIT SAHA is a faculty member in the department of Computer Science and Engineering, NITTTR, Kolkata. He has done his postdoctoral research at the National Research Council, Italy and University of Wroclaw, Poland. He was the visiting research scientist at CWI, Netherlands, INRIA France, IIT-CNR, Italy and ICM in University of Warsaw (UW), Poland. He received his Ph.D. degrees in Computer Science and Engineering and Bioinformatics from Jadavpur University and Polish Academy of Sciences after completing B.Tech and M.Tech. He has co-authored about 100 research papers in various International Journals and Conferences. Dr. Saha is an active member of the board of reviewers in many International Journals. His research interest includes Education Technology, Computational Intelligence, Computational Biology, Machine Learning, Image Processing and Pattern Recognition.

DARIUSZ PLEWCZYNSKI is a professor at University of Warsaw in Center of New Technologies CeNT, Warsaw, Poland. He was postdoc researcher at International Institute for Cell and Molecular Biology in 2001 under the supervision of Dr. Leszek Rychlewski. Dr. Dariusz Plewczynski worked with Dr. Adam Godzik in The Burnham-Sanford Institute in San Diego, CA, USA in 2002. He was a postdoc at Helsinki University, bioinformatics laboratory of Prof. Liisa Holm in 2003. In 2004 he was visiting researcher hosted by Dr Uwe Koch at Merck Research Laboratories (IRBM) in Rome, Italy. From 2002 till 2011, he was the assistant professor at University of Warsaw, Warsaw, Poland and the head of chemo-informatics unit at BioInfoBank Institute, Poznan, Poland. In 2011, He visited Stanford University within Top500 Polish Ministry of Science programme. He received DSc degree (habilitation) in Computer Science and Bioinformatics in 2012 at the Computer Science Institute of Polish Academy of Sciences. His research interests are focused on functional and structural genomics.

***