Comparison of genomes of different species of coronaviruses using spectra of periodicities

L A Miroshnichenko¹, V D Gusev¹ and Yu P Dzhioev²

¹ Sobolev Institute of mathematics SB RAS, Ac. Kaptuga ave., 4, Novosibirsk, 630090, Russia
² Research Institute of Biomedical Technologies of Irkutsk State Medical University, Krasnoe Vostanie str., 1, Irkutsk, 664003, Russia
E-mail: luba@math.nsc.ru, gusev@math.nsc.ru, alanir07@mail.ru

Abstract. In the genomes of different organisms, there are periodicities, i.e. fragments of DNA (RNA)-sequences formed by tandem repetition of the basic monomer (period). The spectra of periodicities with lengths exceeding the ‘noise’ threshold are quite compact and visible even for complete genomes. This makes them an acceptable tool for differentiating closely related objects. The object of analysis in this work are the periodicities at genomes of three species of coronaviruses: MERS, SARS, and SARS-CoV-2. It has been shown that there are markers in the form of periodicities that make it possible to distinguish between these species of coronaviruses. None of the periodicities identified in the genomes of the MERS species (except for the poly-a tract in the 3'UTR) is found in the genomes of SARS and SARS-CoV-2 and vice versa. Revealed periodicities common to SARS and SARS-CoV-2, as well as inherent only to genomes of one species. The number of periodicities in SARS and SARS-CoV-2 significantly exceeds the number of periodicities in random sequences. The periodicities found in almost all genomes of only ‘their’ species are of the greatest interest in terms of revealing the pathogenic potential of the virus.

1. Introduction
Periodicities or short tandem (simple) repeats are widespread molecular markers in genetic and genomic research [1]. They are used for DNA profiling in cancer diagnosis [2], in kinship analysis (especially paternity testing) [3] and in forensic identification [4]. They are also used in genetic linkage analysis to locate a gene or a mutation responsible for a given trait or disease. Microsatellites are also used in population genetics to measure levels of relatedness between subspecies, groups and individuals [5].

The choice of an algorithm for detecting tandem repeats essentially depends on the type of repeats under consideration: perfect [6,7], imperfect [8], or hidden latent [9].

With regard to the problems of classification of DNA (RNA) -sequences, perfect periodicities are of the greatest interest. They are used as molecular markers in the intra- and interspecific classification of organisms [10 – 12].

By the local periodicity P we mean the fragment of a sequence S formed by the tandem k-fold repetition of the basic (minimum possible of the acceptable variants) monomer M (k ≥ 2). For clarity, we will separate the monomers with the symbol ‘-’. For example P = tcaagtcaagtcaag = tcaag-tcaag-tcaag. Here the monomer M = tcaag, its length p = | M | = 5 and the multiplicity k = | P | / p = 3. The
periodicity can be terminated with incomplete monomer, while \( k \) is not an integer. So, for \( P = \text{tcaag-tcaag-tcag-tcag-tc} \), \( k = |P| / p = 17/5 \).

From the entire set of periodicities, we exclude insignificant (‘random’), the length of which does not exceed a given threshold \( r \). The set of periodicities at the sequence \( S \) such that \( |P| \geq r \) will be called the truncated spectrum of periodicities and denoted by \( \mathcal{I}(S) \).

The object of research in this article is the periodicities in the RNA sequences of the genomes of the coronavirus of the three most well-known species: MERS [13], SARS [14] and SARS-CoV-2 [15,16]. The first confirmed case of Middle East respiratory syndrome-related coronavirus (MERS-CoV) was reported in Jeddah, Saudi Arabia in April 2012. By July 2015, MERS-CoV cases had been reported in over 21 countries, in Europe, North America and Asia as well as the Middle East. Severe acute respiratory syndrome coronavirus (SARS-CoV or SARS-CoV-1) caused the 2002–2004 SARS outbreak. Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first known case was identified in Wuhan, China, in December 2019. The disease has since spread worldwide, leading to an ongoing pandemic.

The set of complete genomes of coronaviruses we are investigating (from those presented in the Genbank and Gisaid databases) contains 36 MERS RNA sequences, 24 SARS sequences and 42 SARS-CoV-2: 2 complete genomes for each of 21 countries. A list of genomes is provided in Appendix A, B and C.

Various methods are used to compare the genomes of these species [17, 18]. The purpose of this work is to calculate and analyze the spectra of periodicities for three species of coronavirus, to identify their general and specific features (primarily related to SARS-CoV-2) and to experimentally evaluate the statistical significance of the results. It is assumed that the results obtained will be useful to physicians and biologists involved in a comprehensive study of coronavirus infection using diverse information, in this case, at the genomic level.

2. Description of experiments

To calculate the spectra of periodicities, we used algorithms and programs developed at the Sobolev Institute of Mathematics SB RAS for estimating the complexity of symbolic sequences in the language of multi-type repetitions contained in them [19]. The final product was the decomposition of the sequence into fragments corresponding to one or another type of repeats. The resulting ‘complexity decomposition’ of the sequence explicitly contain fragments formed by tandem repeats.

The algorithm allows detecting all periodicities. Then, of the entire set of periodicities, we leave for consideration only those whose length is not less than a certain threshold value \( r \), i.e. a set \( \mathcal{I}(S) \). With the length of the coronavirus genomes of the order of 30 thousand nucleotides, the choice of the lower value \( r = 12 \) for the lengths of the detected periodicities approximately corresponded to the border between random and nonrandom structures.

The 5'UTR region in the genomes under consideration does not contain periodicities with lengths exceeding the specified threshold. Therefore, for ease of comparison of results, the numbering of positions in the following tables 1-5 coincides with the beginning of the coding region (orf1ab gene). Along with a set of 42 SARS-CoV-2 genomes, which was used for comparison with the SARS and MERS genomes, an expanded set of 560 complete genomes was considered: 320 sequenced in Russia and 240 sequences from other countries (12 strains from 20 countries each). Since the results are consistent, the following are only more visible data based on the results of processing a small set of 42 genomes.

3. Periodicities in the genomes of the MERS species

Table 1 shows the periodicities (\(|P| \geq 12\)) found in almost all genomes of the MERS species (36 RNA sequences from GenBank listed in appendix A). Rare exceptions are indicated in the last column of the table. Positions are given for Middle East respiratory syndrome-related coronavirus isolate 013, complete genome, locus in GenBank MN120514.1, its length is 29840 bp.
In addition, the periodicity tcaact-tcaact whose period is symmetric in position 25488 (orf3) occurs in 23 genomes from of 36.

Table 1. Periodicities found in all genomes of the MERS species.

| No. | Pos. | Region         | Periodicity       | $|P|/p$ | Amino acid sequence with context | Comments |
|-----|------|----------------|-------------------|-------|----------------------------------|----------|
| 1   | 4898 | orf1ab         | ttaggc-ttaggc     | 14/7  | DFTIALIMAYGNCTFGA                 |          |
| 2   | 10840| orf1ab         | aacacc-aacacc-a   | 13/6  | FVMLLVKKHHTFLTLF                  |          |
| 3   | 13076| orf1ab         | attgg-attgg-at    | 12/5  | CNVCQYWIGYGCNCD                   |          |
| 4   | 16780| orf1ab         | cttggca-ctggca    | 12/6  | TVQGPGTGKSHFAI                    |          |
| 5   | 19075| orf1ab         | atggta-atggta     | 12/6  | TPCEVHNGSMIEDI                    | symmetry |
| 6   | 19178| orf1ab         | tacagag-tacagag   | 14/7  | VCRKHATETREYMEAY                  |          |
| 7   | 19363| orf1ab         | cttagt-cttagt     | 12/6  | VEGELPAVNVNKF                    |          |
| 8   | 20910| orf1ab         | aatgca-aatgca     | 12/6  | WTVFCTNASSSEG                  |          |
| 9   | 21076| orf1ab         | aattaa-aattaa-aa  | 14/6  | DLSDKFQLKLGTPVLQ                | symmetry |
| 10  | 28594| gene N         | actgga-actgga     | 12/6  | RWYFYTTGTGPEAAALP                |          |
|     |      | orf8b          |                   |       | GGTTTLELD PKQH                  |          |
| 11  | 29013| gene N         | tgc-tgc-tgc-tgc-t | 13/3  | VITKDDAAAAKNKMRH                |          |

Four of the periodicities listed in table 1 (numbers 2, 5, 9 and 11) have a degenerate alphabet, i.e. do not contain one or two 2 nucleotides.

The sixth column shows the amino acid sequence corresponding to the considered periodicity (in bold) with a left and right context of 6 amino acids. Note that the periodicity at the nucleotide level does not always correspond to the periodicity at the amino acid level.

In 20 from 36 genomes of the MERS species, as well as in 11 from 24 SARS and 13 from 42 SARS-CoV-2, a poly-a tract is found in the 3'UTR region (with different repetition rates). This is the only periodicity that is common for all species of coronaviruses.

Seven from 36 genomes contain periodicities that are not found in other genomes. They are presented in table 2

Table 2. Uncommon periodicities in the genomes of the MERS species.

| No. | Pos. | Region | Periodicity | $|P|/p$ | Genomes in which periodicity occurs |
|-----|------|--------|-------------|-------|-----------------------------------|
| 1   | 3282 | orf1ab | tggtaa-tggta | 12/6  | MW086539 and MW086527             |
| 2   | 18417| orf1ab | gattat-gattat-g| 13/6  | MW086539                          |
| 3   | 21970| gene S | cctc-cctc-cctc| 13/5  | MN120514 and KJ477102             |
| 4   | 25222| gene S | gttcat-gttcat | 12/6  | KJ477102                          |
| 5   | 25486| orf3   | ttcac-ttcac-a | 13/6  | MK967708                          |
| 6   | 28358| gene N | aaacca-aaacca | 12/6  | MT576585                          |
| 7   | 29377| 3'UTR  | a-aagaa-aagaa-a | 12/5  | MH454272 and MN120514, simm.      |

Note that none of the periodicities found in MERS (except for the poly-a tract in 3'UTR) is found in the genomes of SARS and SARS-CoV-2.

The dependence of the number of periodicities on the period length (in total for the entire set of 36 MERS genomes) is shown in figure 1. The maximum (278) falls on the period length 6, which is partly related to the choice of the periodicity length threshold $r = 12$. When this value decreases to $r = 10$, the peak shifts by $p = 5$. It is interesting to note that there are no periodicities with period lengths of 2, 4, 8 and larger.
4. Periodicities in the genomes of species of SARS and SARS-CoV-2

There are eight ‘common’ periodicities ($|P| \geq 12$) in the genomes of the SARS and SARS-CoV-2 coronavirus species. In addition, 17 periodicities are found only in the genomes of the SARS species and 14 only in the genomes of the SARS-CoV-2 species. Tables 3, 4, and 5 include only periodicities that are found in almost all genomes of their class (with a few exceptions). In the comments, the definitions of genomes are replaced by their numbers. The correspondence between numbers and identifiers is shown in Appendix B and C.

Table 3. Periodicities common to genomes of SARS and SARS-CoV-2 coronaviruses.

| No. | Pos. | Region      | Periodicity | $|P|/p$ | Amino acid sequence with context | Comments                  |
|-----|------|-------------|-------------|-------|---------------------------------|--------------------------|
| 1   | 9256 | orf1ab      | tacttg-tacttg | 12/6  | GVYSVFYLYLFYFTN                   | absent at No. 28          |
|     | 9325 |             |             |       | GVYSVFYLYLFYFTN                   |                          |
| 2   | 15506| orf1ab      | gactgagactga | 12/6  | SEAKCWTEDLTGKPH                  |                          |
|     | 15575|             |             |       | SEAKCWTEDLTGKPH                  |                          |
| 3   | 16753| orf1ab      | ctggtactggta | 12/6  | STLQFPGTGKSHFAI                  |                          |
|     | 16822|             |             |       | STLQFPGTGKSHFAI                  |                          |
| 4   | 22447| gene S      | attata-attata-a | 13/6  | TGIADYNYKLPDDFM                  |                          |
|     | 22556|             |             |       | TGIADYNYKLPDDFM                  |                          |
| 5   | 22529| gene S      | taatta-taatta-taa | 15/6  | DATSTGNYNKRYRLR                  | absent at No. 23          |
|     | 22638|             |             |       | DSKVGGNYLYRLFR                   | compl.                   |
| 6   | 25885| gene E      | gtaata-gtaata-g | 15/7  | SEETGLIVNSVLLFL                  |                          |
|     | 26012|             |             |       | SEETGLIVNSVLLFL                  |                          |
| 7   | 27195| orf7a       | tttgct-tttgct-t | 13/6  | TCFSTQFAFACDGRTH                 | absent at No. 8           |
|     | 27315|             |             |       | TCFSTQFAFACDGRTH                 |                          |
| 8   | 28573| gene N      | caa-caa-caa-caa | 12/3  | KVSGKQQQQQQQGQTVTKK              |                          |
|     | 28723|             |             |       | KVSGKQQQQQQQGQTVTKK              |                          |

Positions (first line of each cell) are indicated for SARS coronavirus Tor2, No. 15 in Appendix B and Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2 / human / ITA / INMI1 / 2020, No. 15 in Appendix C (second line). As at table 1 the sixth column shows the amino acid sequence corresponding to the considered periodicity (in bold) with a left and right context of 6-7 amino acids.

When the threshold $r$ is reduced to 11, this list is supplemented with the a-agga-agga-ag periodicity with symmetric period found at positions 14394 (SARS) and 14463 (SARS-CoV-2) and the attac-attaca-a periodicity that terminates gene S (positions 24980 and 25104, respectively).

Table 4 shows the periodicities of length 12 and higher presented in each of the 24 genomes of the SARS species, but not found in the SARS-CoV-2 genomes.
Table 4. Periodicities found in all genomes of the SARS species.

| No. | Pos. | Region   | Periodicity | $|P|/p$ | Amino acid sequence with context | Comments |
|-----|------|----------|-------------|-------|--------------------------------|----------|
| 1   | 838  | orf1ab   | aaga-t-aaga-tca | 13/6  | VFPLNSKVKVIQRPE               |          |
| 2   | 876  | orf1ab   | aaga-aaga-aaga | 12/6  | IQPRVEKKTTEGRM               |          |
| 3   | 1376 | orf1ab   | ttaca-ttaca-ttt | 14/6  | LSERVNNINIVDFHL               |          |
| 4   | 2977 | orf1ab   | ccaaga-ccaga-cc  | 14/6  | EQSEIEPEPEPTPEEP         |          |
| 5   | 7232 | orf1ab   | tctatg-tctatg  | 12/6  | GMKRSFYVYANGGRR               |          |
| 6   | 7924 | orf1ab   | gtgcag-gtgcag  | 14/7  | HHSDEVTGDSCNWFNM                 |          |
| 7   | 9566 | orf1ab   | acaagt-acagnt-a | 13/6  | YLALYNKYYFGSL               |          |
| 8   | 10818| orf1ab   | taagca-taagca  | 12/6  | CAMLLVKHKHAFLCL               |          |
| 9   | 12532| orf1ab   | ggtaca-ggtaca  | 12/6  | RFPKSDGTGITYTELE              |          |
| 10  | 16907| orf1ab   | tggcgc-tggcgc  | 12/6  | CSRIPARAREVECFD               |          |
| 11  | 18845| orf1ab   | tgtt-gttt-gttt  | 13/4  | DKTGDICALWNVCNVD               |          |
| 12  | 22803| gene S   | aatatt-aatatt-aat | 15/6  | IKNQCVNENFENLGTFG                 | symmetry |
| 13  | 22824| gene S   | actggt-actggt  | 12/6  | FNENGFNTGTGVTLPSS              |          |
| 14  | 22966| gene S   | ggtgaa-ggtgaa  | 12/6  | PCAFGGVSIVPUTGN               |          |
| 15  | 26094| -        | tat-tat-tat-tat-t | 13/3  | abesent at No. 9, 17,20         |          |
| 16  | 26206| gene M   | ttctca-ttctca  | 12/6  | WNLQIVGFRLAWIMLL               |          |
| 17  | 26367| gene M   | gattgc-gattgc  | 12/6  | NWVTGGIAIAMACIV               |          |

Positions are indicated for SARS coronavirus Tor2, No. 15 in Appendix B. In addition, in the genome of SARS coronavirus ShanghaiQXC1 (No. 15 in Appendix B) and only in it, the catgaa-catgaa periodicity (12/6) was found at position 700. The nature of the dependence of the number of periodicities on the length of the period (for the entire set of SARS genomes) is similar to that shown in figure 1. The maximum value is 499 at $p = 6$.

Table 5 shows the periodicities of length 12 and higher presented in each of the 42 genomes of the SARS-CoV-2 species that are not found in the SARS genomes.

Table 5. Periodicities found in all genomes of the SARS-CoV-2 species.

| No. | Pos. | Region   | Periodicity | $|P|/p$ | Amino acid sequence with context | Comments |
|-----|------|----------|-------------|-------|--------------------------------|----------|
| 1   | 3816 | orf1ab   | tgacat-tgacat | 12/6  | SATLVSdiditflkk               |          |
| 2   | 4264 | orf1ab   | tttac-tttac-t | 12/6  | VYDGRFMYFTSKTTV               |          |
| 3   | 4423 | orf1ab   | gttct-gttct  | 12/6  | LKVFPATVSVSPDAVT               |          |
| 4   | 8379 | orf1ab   | tctag-tcag-tc  | 12/5  | YLITPVHVMKSHTDF               |          |
| 5   | 10048| orf1ab   | ctaag-ctaag  | 12/6  | SMQNCVNLKKVDTAPN              |          |
| 6   | 10772| orf1ab   | ctttag-cttag-t | 13/6  | TITLSTLVLVQSTQWS              |          |
| 7   | 10911| orf1ab   | tttgct-ttgct  | 12/6  | KHAFLCLFLIPSLAT               |          |
| 8   | 12904| orf1ab   | actggt-actggt | 12/6  | KMLCTHGTGQATITV                |          |
| 9   | 20075| orf1ab   | tga-tga-tga  | 12/6  | IVYGDFSHLQGLGLH               |          |
| 10  | 20314| orf1ab   | tgctag-tgctag | 14/7  | SKVVKVTIDYTEISFM               |          |
| 11  | 20559| orf1ab   | acata-acata  | 12/6  | LCYQNLTLTAVPYNM               |          |
| 12  | 23319| gene S   | tcagc-tcagc-t | 13/6  | GICASYQTOINSPRRRA              |          |
| 13  | 25228| orf3a    | cgata-cgatac  | 12/6  | VRATTATPIQASLPFG                |          |
| 14  | 25504| orf3a    | ggcctt-ggcctt  | 12/6  | VRLLMRLCWRKCRS               |          |

Positions are indicated for the genome of Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2 / human / ITA / INMI1 / 2020, No. 15 in Appendix C.
In addition to the periodicities presented in table 5, no other periodicity was revealed that is found in individual genomes of the set. In an expanded set of 560 SARS-CoV-2 genomes, only 3 strains were identified, in which periodicities not indicated in Tables 3 and 5 are presented:

Poland/PL_P5/2020; поз. 25160 : aacttt-aacttt (12/ 6);
EPI_ISL_450249#Russia#Saint-Petersburg#2020-04-21; поз. 25104 : attac-attac-at (12/ 5) and
EPI_ISL_450251#Russia#Saint-Petersburg#2020-04-20; поз. 18995: taacct-taacct (12/ 6).

5. Comparison with pseudo-random sequences

The data for the experiment was generated by randomly shuffling elements of each sequence from the original sets. 100 random copies were obtained for each genome. Thus, 3600 analogs of the genomes of the MERS species were obtained, 2400 – of the genomes of the SARS species, and 4200 for SARS-CoV-2. The results of the experiment are presented in table 6. For the number of periodicities, their minimum, average and maximum values in pseudo-random sequences are indicated. The values of the absolute maximum for the entire set of sequences and the average value of the lengths of the maximum periodicities for each of the sequences are given. The last column of table 6 shows the percentage of nucleotides in the sequences and their periodic fragments. Here \( \pi(x) \) the percentage of nucleotide \( x \) (\( x \in \{a, c, g, t\} \)) in the periodicities found in random sequences, \( \pi_a(x) \) the percentage of nucleotide \( x \) in the original genomes (in shuffled sequences, this value is preserved ), \( \pi_p \) – in the periodicities of the original genomes, \( \pi_a \) – in the periodicities of the original genomes without taking into account the poly-a tract.

### Table 6. Periodicities in random analogs of the original selection

| Species of coronavirus | Number of periodicities | Max. and average length | Percentage of nucleotides in sequences and their periodic fragments |
|------------------------|-------------------------|-------------------------|---------------------------------------------------------------------|
|                        | min | max | average | max | average | a | t | g | c |
| MERS                   | 2   | 25  | 11.34   | 27  | 15.8    | \( \pi_r \) | 26.56 | 41.15 | 17.05 | 15.23 |
|                        |     |     |         |     |         | \( \pi_s \) | 26.25 | 32.59 | 20.92 | 20.24 |
|                        |     |     |         |     |         | \( \pi_p \) | 37.61 | 25.07 | 21.51 | 15.82 |
|                        |     |     |         |     |         | \( \pi_a \) | 34.29 | 26.40 | 22.65 | 16.66 |
| SARS                   | 1   | 23  | 10.82   | 26  | 15.7    | \( \pi_r \) | 31.38 | 36.77 | 16.84 | 15.01 |
|                        |     |     |         |     |         | \( \pi_s \) | 28.49 | 30.75 | 20.82 | 19.94 |
|                        |     |     |         |     |         | \( \pi_p \) | 37.86 | 31.84 | 17.56 | 12.75 |
|                        |     |     |         |     |         | \( \pi_a \) | 35.98 | 32.80 | 18.09 | 13.13 |
| SARS-CoV-2             | 2   | 27  | 12.65   | 28  | 16.1    | \( \pi_r \) | 34.29 | 39.09 | 14.53 | 12.09 |
|                        |     |     |         |     |         | \( \pi_s \) | 29.93 | 32.14 | 19.61 | 18.33 |
|                        |     |     |         |     |         | \( \pi_p \) | 29.13 | 40.91 | 15.52 | 14.44 |
|                        |     |     |         |     |         | \( \pi_a \) | 26.67 | 42.33 | 16.06 | 14.94 |

Note that the maximum length of periodicities in most of the original sequences is 14 for MERS and 15 for SARS and SARS-CoV-2. In some random sequences, periodicities of significantly longer lengths can be generated randomly (26 – 28), however, the average values of the maximum lengths (15.7 – 16.1) for the entire set of random texts are comparable to the values for the original genomes.

The number of periodicities in most MERS genomes is 11 (table 1), which differs little from the average number of periodicities in random sequences. In the genomes of SARS, however, the number of periodicities is already 25 – 27, which is significantly higher than the average number for random sequences and even higher than the maximum value. The number of periodicities in the genomes SARS-CoV-2 (22 – 24) is slightly less than the maximum value in random sequences, but significantly (almost 2 times) exceeds the average. This may indicate that the periodicities presented in the genomes of SARS have a significant functional capacity. This issue requires a separate study.

The dependence of the number of periodicities on the length of the period both in real genomes and in their random analogs is shown in table 7. Note again that no periodicities with a monomer length of...
2 or more than 8 were found in any of the coronavirus genomes, but periodicities with a period length of 4 were revealed in the SARS genomes. The maximum number of periodicities is detected at a period length of 6. With a decrease in the threshold to \( r = 10 \), the number of periodicities increases approximately 3 times in the genomes of coronaviruses and 4 times in their random counterparts, and the maximum number of periodicities is reached with a period length of 5.

**Table 7.** Dependence of the number of periodicities on the length of the period (at %)

| \( p \) | MERS  | MERS (random) | SARS  | SARS (random) | SARS-CoV-2  | SARS-CoV-2 (random) |
|-------|-------|---------------|-------|---------------|-------------|---------------------|
| 1     | 4.49  | 0.31          | 1.75  | 0.18          | 1.60        | 0.27                |
| 2     | 0.23  |               | 0.25  |               |             | 0.37                |
| 3     | 8.09  | 1.05          | 7.62  | 1.06          | 4.49        | 1.28                |
| 4     | 4.03  | 3.81          | 3.96  |               |             | 4.19                |
| 5     | 8.99  | 16.58         | 16.02 | 4.27          | 16.58       |                     |
| 6     | 62.47 | 57.60         | 79.21 | 57.43         | 80.66       | 56.40               |
| 7     | 15.96 | 14.95         | 7.62  | 15.62         | 8.97        | 15.40               |
| 8     | 3.88  |               | 4.02  |               |             | 4.06                |
| 9     | 1.02  |               | 1.10  |               |             | 1.09                |
| 10    | 0.26  |               | 0.29  |               |             | 0.27                |
| 11    | 0.08  |               | 0.05  |               |             | 0.07                |
| 12    | 0.02  |               | 0.01  |               |             | 0.02                |
| 14    | 0.005 |               | 0.01  |               |             | 0.004               |

6. Conclusion

The analysis of a limited set of genomes of three species of coronaviruses (MERS, SARS and SARS-CoV-2) shows that:

- There are markers for the differentiation of coronavirus species at the level of RNA sequences. As such, tandem repeating chains of symbols (periodicities) can be used.
- The number of periodicities with a length exceeding the threshold value \( r = 12 \) in genomes of the MERS species is relatively small and differs little from ‘random’ (shuffled) sequences of the same length and with the same composition of elements. However, in the genomes of the SARS and SARS-CoV-2 species, this number is more than twice as high.
- None of the periodicities identified in the genomes of the MERS species (except for the poly-a tract in the 3’UTR) is found in the genomes of SARS and SARS-CoV-2 and vice versa.
- The number of periodicities common to all genomes of the SARS and SARS-CoV-2 species is approximately 2 times less than the number of periodicities characteristic of only one of the species. It is the periodicities found in almost all genomes of only ‘their’ species that are of greatest interest in terms of identifying the pathogenic potential of the virus.
- In view of the limitedness of the initial set of genomes, the formulated conclusions are preliminary in nature and are subject to verification and refinement on a more voluminous material.
Acknowledgments
The study was carried out within the framework of the state contract of the Sobolev Institute of Mathematics (project no. 0314-2019-0015).

Appendix A. List of MERS coronavirus strains (GenBank)
1. JX869059.2 Human betacoronavirus 2c EMC/2012
2. KJ614529.1 Human betacoronavirus 2c Jordan-N3/2012 isolate MG167
3. NC_038294.1 Betacoronavirus England 1 isolate H123990006
4. MH454272.1 Middle East respiratory syndrome-related coronavirus strain HCoV-EMC
5. MK967708.1 Middle East respiratory syndrome-related coronavirus isolate Merscov/Egypt/Camel/AHRI-FAO-1/2018
6. MN120514.1 Middle East respiratory syndrome-related coronavirus isolate 013
7. MW086539.1 Middle East respiratory syndrome-related coronavirus isolate MERS-CoV/JC33/Ramtha
8. MW086527.1 Middle East respiratory syndrome-related coronavirus isolate MERS-CoV/JC17/Azraq
9. KU740200.1 Middle East respiratory syndrome-related coronavirus isolate MERS-CoV/camel/Egypt/NRCE-NC163/2014, partial genome
10. MT576585.1 Middle East respiratory syndrome-related coronavirus isolate MA_30
11. KT029139.1 Middle East respiratory syndrome coronavirus isolate MERS-CoV/KOR/KNIH/002_05_2015
12. KT006149.2 Middle East respiratory syndrome coronavirus strain ChinaGD01
13. MF988722.1 Middle East respiratory syndrome-related coronavirus strain camel/UAE_415915_W6_2015
14. KJ556336.1 Middle East respiratory syndrome coronavirus isolate Jeddah_1_2013
15. KJ477102.1 Middle East respiratory syndrome coronavirus
16. KJ829365.1 Middle East respiratory syndrome coronavirus isolate Florida/USA-2_Saudi Arabia_2014
17. KJ361503.1 Middle East respiratory syndrome coronavirus isolate Hu/France-FRA2_130569-2013_Isolate_Sanger
18. KF192507.1 Middle East respiratory syndrome coronavirus
19. KJ156952.1 Middle East respiratory syndrome coronavirus isolate Riyadh_4_2013
20. KJ156949.1 Middle East respiratory syndrome coronavirus isolate Taif_1_2013
21. KJ156910.1 Middle East respiratory syndrome coronavirus isolate Hafr-Al-Batin_2_2013
22. KF745068.1 Middle East respiratory syndrome coronavirus isolate FRA/UAE
23. KF600630.1 Middle East respiratory syndrome coronavirus isolate Buraidah_1_2013
24. KF600620.1 Middle East respiratory syndrome coronavirus isolate Bisha_1_2012
25. KX034100.1 Middle East respiratory syndrome coronavirus isolate MERS-CoV/KOR/Seoul/177-3-2015
26. KP209313.1 Middle East respiratory syndrome coronavirus strain Abu Dhabi_UAE_26_2014
27. KU308549.1 Middle East respiratory syndrome coronavirus isolate Korea/Seoul/SNU1-035/2015
28. KJ813439.1 Middle East respiratory syndrome coronavirus strain IndianA/USA-1_Saudi Arabia_2014
29. KF186565.1 Middle East respiratory syndrome coronavirus isolate Hu/Taif/KSA-7032/2014/S530del
30. KF961222.1 Middle East respiratory syndrome coronavirus isolate Qatar4
31. KT225476.2 Middle East respiratory syndrome coronavirus isolate MERS-CoV/THA/CU/17_06_2015
32. KX154694.1 Middle East respiratory syndrome coronavirus isolate Hu/Artawiyah-KSA-13328/2016
33. KT861627.1 Middle East respiratory syndrome coronavirus isolate Hu/Jordan-2014001068_2014
34. KU710265.1 Middle East respiratory syndrome coronavirus isolate Hu/Tai/KSA-7032/2014/S530del
35. KU710264.1 Middle East respiratory syndrome coronavirus isolate Hu/Tai/KSA-7032/2014
36. KT806054.1 Middle East respiratory syndrome coronavirus isolate Hu/Najran-KSA-C20915/2015

Appendix B. List of SARS coronavirus strains (GenBank)
1. JN854286.1 SARS coronavirus HKU-39849 isolate recSARS-CoV HKU-39849
2. DQ182595.1 SARS coronavirus ZJ0301 from China
3. AY394850.2 SARS coronavirus WHU
4. FJ882960.1 SARS coronavirus ExoN1 isolate P3pp34
5. FJ882949.1 SARS coronavirus wtic-MB isolate P3pp23
6. AY279354.2 SARS coronavirus BJ04
7. AY278489.2 SARS coronavirus GD01
8. AY772062.1 SARS coronavirus WH20
9. AY348314.1 SARS coronavirus Taiwan TC3
10. AY278741.1 SARS coronavirus Urbani
11. DQ898174.1 SARS coronavirus strain CV7  
12. AY310120.1 SARS coronavirus FRA  
13. AY291451.1 SARS coronavirus TW1  
14. AY283798.2 SARS coronavirus Sin2774  
15. NC_004718.3 SARS coronavirus Tor2  
16. AY283797.1 SARS coronavirus Sin2748  
17. AP006561.1 SARS coronavirus TWY genomic RNA  
18. AY508724.1 SARS coronavirus NS1  
19. AY395003.1 SARS coronavirus ZS-C  
20. AY394998.1 SARS coronavirus LC1  
21. AY394993.1 SARS coronavirus HGZ8L2  
22. AY394986.1 SARS coronavirus HSZ-Cb  
23. AY394979.1 SARS coronavirus GZ-C  
24. AY463059.1 SARS coronavirus ShanghaiQXC1

Appendix C. List of SARS-CoV-2 coronavirus strains (Gisaid)  
1. EPI_ISL_428906#Russia#Republic_of_Buryatia#2020-03-27  
2. EPI_ISL_428891#Russia#Republic_of_Sakha#2020-03-25  
3. Nan chang/JXN3T4/2020  
4. Shanghai/SH0007/2020  
5. Egypt/CUNCI-HGC3I06/2020  
6. Egypt/CUNCI-HGC3I07/2020  
7. France/HF3677/2020  
8. France/ARA12877/2020  
9. Germany/NRW-MPP-1/2020  
10. Germany/NRW-MPP-33/2020  
11. Greece/207_33548/2020  
12. Greece/134/2020  
13. Israel/CVL-n-6112/2020  
14. Israel/n14285/2020  
15. Italy/INMI16/2020  
16. Italy/IZSM-32798-45/2020  
17. Japan/PG-0200/2020  
18. Japan/PG-0207/2020  
19. Kazakhstan/26474/2020  
20. Kazakhstan/26478/2020  
21. Poland/PL_P21/2020  
22. Poland/PL_P23/2020  
23. South Korea/KCDC2077/2020  
24. South Korea/KCDC2078/2020  
25. Spain/NC-001398/2020  
26. Spain/Zaragoza2422/2020  
27. Thailand/NIH-15/2020  
28. Thailand/NIH-2296/2020  
29. Turkey/KU-037/2020  
30. Turkey/HSGM-10232/2020  
31. Ukraine/203100331/2020  
32. Ukraine/203100333/2020  
33. England/NORT-2846F1/2020  
34. England/NORT-292177/2020  
35. USA/NY-PV08486/2020  
36. USA/CA-ALSR-0943-SAN/2020  
37. Brazil/PR-LRV-01/2020  
38. Brazil/RJ-1402/2020  
39. Mexico/QuintanaRoo-InDRE-14/2020  
40. Mexico/CDMX-INCMNSZ_04/2020
References
[1] Li Y C, Korol A B, Fahima T, Beiles A and Nevo E 2002 Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review Molecular Ecology 11 (12) pp 2453–65
[2] Sideris M and Papagrigoriadis S 2014 Molecular biomarkers and classification models in the evaluation of the prognosis of colorectal cancer Anticancer Research. 34 (5) pp 2061–8
[3] Lászik A, Brinkmann B, Sótonyi P and Falus A 2000 Automated fluorescent detection of a 10 loci multiplex for paternity testing Acta Biologica Hungarica. 51 (1) pp 99–105
[4] Butler J M 2005 Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers (New York: Elsevier Academic Press)
[5] Manel S, Schwartz M K, Luikart G and Taberlet P 2003 Landscape genetics: combining landscape ecology and population genetics. Trends in Ecology & Evolution. 18 (4) pp 189–97
[6] Crochemore M and Rytter W 1996 Text Algorithms (New York: Oxford University Press) p 412
[7] Gusev V D, Miroshnichenko L A and Chuzhanova N A 2009 Identification of fractal-like structures in DNA sequences Information Science & Computing. International Book Series 8: Classification, Forecasting, Data Mining (ITHEA, Sofia) pp 117–23
[8] Benson G 1999 Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Research 27(2) pp 573–80
[9] Chaley M B, Tyulko Zh S and Kutyrykin V A 2020 Coding Structure for the ORF1ab, S, M and N Coronavirus Genes Mathematical Biology and Bioinformatics 15 (2) pp 441–54
[10] Gusev V D, Miroshnichenko L A, Titkova T N, Dzhioev Y P, Kozlova I V and Paramonov A I 2018 Structured RNA markers for genotyping of tick-borne encephalitis virus Mathematical Biology and Bioinformatics 13 (1) pp 13–37
[11] Miroshnichenko L A and Gusev V D 2021 Complete spectra of periodicities in the problems of differentiation of closely related bacterial genomes J. Phys.: Conf. Ser. 1715 012026
[12] Gusev V D, Miroshnichenko L A and Dzhioev Yu P 2019 Fractal characteristics of genomes as an instrument of intra- and interspecific classification of organisms Proc. Int. Conf. Advanced Mathematics, Computations and Applications Mathematics (AMCA-2019) within Marchuk Scientific Readings (Novosibirsk) pp 122–6
[13] Lu R et al 2015 Complete genome sequence of Middle East respiratory syndrome coronavirus (MERS-CoV) from the first imported MERS-CoV case in China. Genome Announce 3 (4) e00818-15
[14] Qin E’de et al 2003 A complete sequence and comparative analysis of a SARS-associated virus (isolate BJ01). Chin Sci Bull 48 (10) pp 941–8
[15] Wu F at al 2020 Complete genome characterisation of a novel coronavirus associated with severe human respiratory disease in Wuhan, China. bioRxiv doi:10.1101/2020.01.24.919183
[16] Wu A et al 2020 Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China Cell host & Microbe 27 (3) pp 325–8
[17] Rehman H A, Ramzan F, Basharat Z, Shakeel M, Khan M U G and Khan I A 2021 Comprehensive comparative genomic and microsatellite analysis of SARS, MERS, BAT-SARS and COVID-19 Coronaviruses. Journal of Medical Virology 93 (7) pp 4382–91
[18] Maldonado L, Bertelli A and Kamenetzky L 2021 Molecular features similarities between SARS-CoV-2, SARS, MERS and key human genes could favour the viral infections and trigger collateral effects Scientific Reports 11 (1) 4108
[19] Gusev V D, Nemytikova L A and Chuzhanova N A 1999 On the complexity measures of genetic sequences Bioinformatics 15 (12) pp 994–9