Mutation hotspots of SARS-CoV-2 RNA motifs conserved in betacoronaviruses

N S Kobalo¹, A A Kulikov¹ and I I Titov²

¹Institute of Computational Mathematics and Mathematical Geophysics SB RAS, 6, Ac. Lavrentieva ave., Novosibirsk, 630090, Russia
²Institute of Cytology and Genetics SB RAS, 10, Prospekt Lavrentyeva, Novosibirsk, 630090, Russia

E-mail: rerf2010rerf@yandex.ru

Abstract. The pandemic of the coronavirus infection COVID-19, which began at the end of 2019 and caused by the SARS-CoV-2 virus, has led to unprecedented consequences in the world. By the end of May 2021, in the world there were 167 million infected and 3.5 million died directly from infection [1]. SARS-CoV-2 is a beta coronavirus, so it shares many conserved fragments with other known viruses of this type [2]. Since the beginning of the spread of the COVID-19, one of the important issues of research of the SARS-CoV-2 virus has been the search for its conserved RNA motifs and their functional annotation. These motifs are potential targets for the treatment and diagnosis of a disease caused by the virus. This report examines the structural RNA fragments of SARS-CoV-2, similar to the corresponding fragments in other beta coronaviruses [2]. For these RNA motifs the nucleotide variability during the spread of the virus, depending on their secondary structure, was investigated. All the motifs display the similar background variability although contain hypervariable positions.

1. Introduction

The SARS-CoV-2 virus, which is the cause of the COVID-19 pandemic, is a type of RNA virus. Its genome contains about 30 thousand nucleotide bases. Different fragments of the genome are required to encode different types of proteins needed to maintain the viral propagation mechanism. Also, the genome contains cis-acting RNAs that do not encode proteins but are also necessary for viral replication [2]. These RNAs have a conserved structure that supports their functionality. These two properties - importance for the spread of the virus and sequence conservatism - make these fragments good targets for diagnostic methods and antiviral drugs. Over the past year and a half, a huge amount of data on SARS-CoV-2 has been accumulated, including the hundreds of thousands of high-quality sequenced genomes from all over the world, allowing us to trace its variability. All this opens up wide opportunities for bioinformatics study of the virus.

SARS-CoV-2 is a beta-coronavirus, and it shares many conserved fragments with other viruses of this type. In [2], a set of such fragments was presented, and models of their structure were proposed. In this work, we used these results and investigated the features of the variability of some conserved fragments.
2. General characteristics
SARS-CoV-2 is a positive-sense, single-stranded RNA coronavirus. The genome of SARS-CoV-2 is about 30,000 nucleotides long. It contains 13-15 open reading frames (ORFs), 12 of which are functional. GC bases constitute 38% of the genome. There are 11 protein-coding genes and 12 expressed proteins. The genome structure is in many ways similar to SARS-CoV and MERS-CoV, they have about 82% identity. At the same time, the identity of the basic structural proteins is more than 90%.

The main structural proteins of the virus are E (envelope protein), M (membrane protein), N (nucleocapsid protein), S (thorn glycoprotein). Spike glycoprotein plays an important role in the spread of the virus because it is responsible for the binding of the virus to the host cell during infection. Therefore, this protein is one of the targets for antiviral vaccines. It has been shown that mutations are concentrated in the S, N, and ORF3a genes. Higher mutational activity in the structural proteins of the virus is needed to escape the immune response and adapt to the host.

In addition to the coding regions of RNA, non-coding regions are also important for the life cycle of the virus. Non-coding fragments of the genome that have a conserved secondary structure of RNA are called structural. For example, the structures of the 5' and 3' untranslated regions (UTRs) are well studied and it was shown that similar structures are also present in the SARS-CoV-2 related coronaviruses. Other conserved structural RNA regions are less well understood. In work [2], 106 structural regions were presented and their models of the secondary structure were proposed.

UTR fragments located at the 5' and 3' ends of the coronavirus genome, respectively, are examples of functional RNA structures. The 5'-UTR consists of 5 domains SL1-SL5. Domains SL1 and SL2 are conserved in beta-coronaviruses. The SL5 domain is known to be responsible for subgenomic RNA synthesis, while the SL5 domain is involved in packaging. The mutational activity of the SL5 site was, among others, considered in this work. The 3'-UTR also contains domains that play an important role in the regulation of viral RNA synthesis and, probably, translation.

The general structure of the SARS-CoV-2 genome is shown in figure 1.

![Figure 1](image)

**Figure 1.** Genome structure of CARS-CoV-2. (a) Positions of ORFs and main structural proteins. (b) Positions of the 106 structural RNA fragments.

3. Materials and methods

3.1. Datasets
The dataset of the conserved structural fragments of the virus considered in this work, as well as their secondary structures, was taken from [2].

The SARS-CoV-2 WIV04 reference was downloaded from GISAID [6].

Mutational activity was calculated based on the mutation annotated tree UShER [7]. The tree contains 1139190 genomes sequenced from 2019-12-01 to 2021-07-08.
3.2. Frequency of Mutation
As a source of data on mutations, we took a tree prepared by the UShER project [7]. This project provides a pre-calculated phylogenetic tree built based on sequenced sequences available in open sources: GenBank, COG-UK, and The China National Center for Bioinformation.

In the UShER phylogenetic tree, each node is annotated with all mutations that occurred during the transition from the ancestor to this node (figure 2). The mutation frequency in the position we are considering is equal to the ratio of the number of mutations this position to the total number of sequences in the tree. We were interested in mutation frequency in general. However, by normalizing the number of events in a position by the length of the edges of the tree, which is equal to the number of mutations in a sequence belonging to a given node, one can also calculate the mutation rate.

Figure 2. An example of the UShER tree. Each node of the tree is provided with a set of annotations containing information about mutations. For example, in this picture, during the transition from the ancestor to the node S1, the nucleotide C was replaced by A at position 1.

With the mutation frequency calculated, it is possible to estimate the stability of the secondary structure of the desired fragment. For this, the difference between the mean frequency of mutations in free and paired nucleotides was calculated. For further analysis, we took two fragments from different ends of the resulting distributions – one with a predominance of mutations in loops and second in stems.

4. General characteristics of variability
The analysis included 106 conserved fragments of SARS-CoV-2 of length 120nt, presented in [2]. In addition to the fragments themselves, their secondary structure is presented in this work. All nucleotide positions in each of the fragments were divided into 2 groups according to their belonging to the stems (paired bases) or loops (free bases) of the corresponding secondary structure. Further, for each of these two groups, the average mutation frequency for all positions in the group was calculated.

As a general characteristic of the evolutionary stability of the structure, the difference between the mean frequency of mutation in the loop and stem was calculated.

On average, for fragments, the frequency of mutation in loops is higher than in stems. This means that free nucleotides mutate more often than paired ones, i.e. the secondary structure tends to persist. However, the inverse case does not necessarily mean the opposite as one can see below.

Let us consider two fragments, 16 and 75, as an example: the first has a maximum positive frequency difference, and the second has a negative one (table 1).
Table 1. Fragment characteristics.

| Number of fragment [2] | Position in the refseq NC_045512.2 | Description of the genome region |
|------------------------|-------------------------------------|----------------------------------|
| 16                     | 158-277                             | SL5 domain of UTR-5’              |
| 75                     | 11730-11849                         | Gene: ORF1ab, from 758 position of nsp6 non-structural protein, genbank YP_009724389 |
| 96                     | 26608-26727                         | Gene: from 86 position of M membrane glycoprotein |

Fragment #16 is the SL5 domain of UTR-5’ of the virus fragment, which plays an important role in the packaging of the virus [8].

Fragment #75 belongs to the gene encoding the nsp6 protein. This fragment contains mutations in paired nucleotides, but these mutations preserve the structure. The Figure 3 shows the structure of the fragments and marks the positions with the highest frequency of mutation. Nucleotide substitutions occurring at these positions preserve paired bases as shown in Table 2. This table shows that mutations do not destroy the structure, even though they are in the stems. For example, a C-G pair in the positions 16-101 can mutate into a U-G pair. Two of four mutations (positions 17 and 20) change the encoded amino acids what suggests that conservation of RNA secondary structure is more important.

Table 2. Mutational hotspots in the fragment 75. In the column "position" the position number is indicated. "Complementary base" shows the base with which the base in question forms a pair. "Substitution" – shown are two most frequent mutations at the position.

| Position in the fragment | Complementary base | Substitution |
|-------------------------|--------------------|--------------|
| 17                      | G                  | C->U(97%), U->C(1%)          |
| 20                      | G                  | C->U(96%), U->C(2%)          |
| 28                      | G                  | C->U(94%), U->C(0.2%)        |
| 52                      | U                  | A->G(97%), G->A(2%)          |
| 94                      | Not paired         | C->U(96%), U->C(2%)          |
Figure 3. Structural fragment #16 (a) and #75 (b). Positions with a high frequency of mutation are circled in bold. In the fragment 16, looped nucleotides have a high frequency of mutations, which confirms the evolutionary conservatism of this secondary structure. On the contrary, in the fragment 75, paired nucleotides at positions 17, 20, 28, 52, 94 have a higher frequency of substitutions. However, these substitutions, as shown above, do not destroy the secondary structure. (c) Nucleotide and corresponding protein sequence of the region of fragment 75 with mutating bases. Mutations at positions 17 and 20 (marked in gray) change the encoded aminoacid, while at positions 28, 52 and 94 they do not.

5. Mutation hotspot
Additionally, we determined the positions with the maximum mutation frequency for all the fragments considered. One of the positions that have the maximum frequency is the already discussed position 20 of fragment 75. The second is the position 73 of the fragment #96 (genbank YP_009724393), shown in figure 4.

Fragment #96 belongs to the M gene encoding a membrane glycoprotein [9]. Position in the reference sequence NC_045512.2 - 26608-26727. The basic C nucleotide from the reference sequence at position
73 tends to be replaced by the U nucleotide in the third codon position and thus retains the phenylalanine in the encoded protein.

(a) Nucleotide and corresponding protein sequence of the region of the fragment with mutating bases. Nucleotide C at the position 73 can mutate into U, which preserves the amino acid F. The mutations in the positions 37 and 82 also preserve the encoded amino acids, but 14 and 17 do not.

(b) The position 73 has one of the highest mutation frequency among all positions of all considered domains is circled in bold. Other hotspots with lower mutation rates are circled in gray.

6. Conclusions
In this work, the mutational activity of the conserved structural fragments of the SARS-CoV-2 coronavirus was examined. Three fragments were examined in detail.

It has been shown that the SL5 domain of 5'-UTR has a secondary structure that is conserved during evolution. The structural domain of the nsp6 gene has a similar property. For this fragment, the preservation of the secondary structure may be more important than the preservation of the protein sequence.

In addition, figure 5 shows the distribution of the mutation frequency depending on the type of position. The mean mutation frequency for the fragment 16 is 0.000048, for the 75 - 0.000031, and for the 96 - 0.000023. We considered three types of positions - a free nucleotide in the secondary structure
(loop), an inner nucleotide of a paired strand (stem), and a nucleotide at the edge at the helix. For the fragment 16, mutations leave the secondary structure stable. Therefore, in the right part of the histogram, in the region of high mutation frequencies, a large frequency in the loops is noticeable. At the edges of the strands, mutations are also strong, and in strands they are weakest. The fragment 96 is also rather under stabilizing selection. In the fragment 75, in the high-frequency region, the stems, loops, and edges are almost equal.

![Figure 5](image)

**Figure 5.** Distribution of mutation frequencies depending on the type of position. “Loop”: for the unpaired nucleotides. “Stem” for the internal nucleotides of the paired helix. “Edge” for the nucleotides of the helix at the helix-loop border. The x-axis shows the mutation frequency, and the y-axis is the probability of a given type of position with a given frequency.

(a) Fragment 16. (b) Fragment 75. (c) Fragment 96.

**References**

[1] Roser M, Ritchie H, Ortiz-Ospina E and Hasell J 2020 Coronavirus pandemic (COVID-19) https://ourworldindata.org/coronavirus

[2] Rangan R, Zheludev I N, Hagey R J, et al 2020 RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look *RNA* 26 937–59

[3] Naqvi A, Fatima K, Mohammad T, Fatima U, Singh I K, Singh A, Atif S M, Hariprasad G, Hasan G M and Hassan M I 2020 Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: structural genomics approach *Biochim Biophys Acta Mol Basis Dis.* 1866(10):165878

[4] Logunov D Y, Dolzhikova I V, Zubkova O V, Tukhvatulin A I, Shcheblyakov D V, Dzharullaeva
A S, Grousova D M, Erokhova A S, Kovyrshina A V, Botikov A G, et al 2020 Safety and immunogenicity of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine in two formulations: two open, non-randomised phase 1/2 studies from Russia *Lancet* 396(10255) 887–97

[5] Durmaz B, Abdulmajed O and Durmaz R 2020 Mutations observed in the SARS-CoV-2 spike glycoprotein and their effects in the interaction of virus with ACE-2 receptor *Medeni Med J.* 35(3) 253–60

[6] Shu Y and McCauley J 2017 GISAID: global initiative on sharing all influenza data–from vision to reality. *Euro Surveill.* 22 30494

[7] Turakhia Y, Thornlow B, Hinrichs A S, et al 2021 Ultrafast sample placement on existing trees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic *Nat. Genet.* 53 809–16

[8] Chen S and Olsthoorn R 2010 Group-specific structural features of the 5′-proximal sequences of coronavirus genomic RNAs *Virology* 401(1) 29-41

[9] Davidson A, Williamson M, Lewis S, Shoemark D, Carroll M, Heesom K, Zambon M, Ellis J, Lewis P, et al 2020 Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein *Genome Med.* 12(1) 68