Putative Secondary Structure at 5'UTR as a Potential Antiviral Target against SARS-CoV-2

SARS-CoV-2 is an enveloped positive-sense single-stranded RNA coronavirus that causes COVID-19, of which the current outbreak has resulted in a high number of cases and fatalities throughout the world, even vaccine doses are being administered. The aim of this work was to scan the SARS-CoV-2 genome in search for therapeutic targets. We found a sequence in the 5'UTR (NC\_045512:74-130), consisting of a typical heptamer next to a structured region that may cause ribosomal frameshifting. The potential biological value of this region is relevant through its low similarity with other viruses, including coronaviruses related to SARS-CoV, and its high sequence conservation within multiple SARS-CoV-2 isolates. We have predicted the secondary structure of the region by means of different bioinformatic tools. We have suggested a most probable secondary structure to proceed with a 3D reconstruction of the structured segment. Finally, we carried out virtual docking on the 3D structure to look for a binding site and then for drug ligands from a database of lead compounds. Several molecules that could be probably administered as oral drugs show promising binding affinity within the structured region, and so it could be possible to interfere its potential regulatory role.

Keywords: SARS-CoV-2; frameshifting; 5'UTR; pseudoknot; molecular docking.

INTRODUCTION

On March 11th, the World Health Organization (WHO) declared COVID-19 a clinical pandemic (primarily pneumonia and gastroenteritis) caused by the SARS-CoV-2 virus. As of end October 2021 the pandemic outbreak has caused almost five million deaths worldwide, although almost 7 million peo-
Illumina, and they demonstrated frameshifting. This area.

virus may produce a fusion protein that overlaps the regions.

frameshift in the SARS-CoV genome. By this mechanism, the

frameshifting signal has two essential elements: a characteri-

tic heptanucleotide called the ‘slippery’ sequence, at which the ribosome-bound tRNAs slip into the −1 frame, and an adjacent mRNA secondary structure that stimulates this slippage process. The intermediate sequence between these two elements also has a typical size of less than twelve nucleotides. Often the secondary structure is more complex than a simple stem-loop between palindromic sequences, expanding into pseudoknots [5]. In terms of structure, a pseudoknot forms upon the base-pairing of a single-stranded region of RNA in the loop of a hairpin to a stretch of complementary nucleotides elsewhere in the RNA chain.

A set of bioinformatic tools has already been developed to predict these structures [6]. The mechanism of action of pseudoknots is not completely understood; some authors suggest that it appears to be linked to the helicase activity of the ribosome. When pseudoknots are located in coding regions, they modulate the elongation and termination steps of translation: the ribosome is able to switch from the zero reading frame to the −1 frame and translation continues in the new frame. When pseudoknots are in non-coding regions, they act on the regulation of the initiation of protein synthesis and on template recognition by the viral replicase guiding viral replication and packaging [7].

All coronaviruses have been reported to utilize programmed −1 ribosomal frameshifting to control the expression of their proteins. In 2005, Plant et al. [8] identified a three-stemmed mRNA pseudoknot inducing an efficient −1 ribosomal frameshift in the SARS-CoV genome. By this mechanism, the virus may produce a fusion protein that overlaps the regions ORF1a and ORF1b. This element encodes an ORF1ab polyprotein involved in ablating the host cellular innate immune response. Mutations affecting this structure decreased the rates of −1 PRF and had deleterious effects on the virus propagation. Recently, Kelly et al. [9] described the same pseudoknot in SARS-CoV-2, and they demonstrated frameshifting. This area

is highly conserved between SARS-CoV and SARS-CoV-2, as there is only one single nucleotide difference, a C to A substitution at position 13,533 bp.

The frameshifting regions could be used as a target to fight viral infection [9]. Starting with early studies, point mutations at the slippery sequence have proved to have an important effect on viral replication [8]; thus, they can be also interesting points in the engineering of an attenuated virus for vaccine development. The inhibition of these regions by peptide antisense oligomers was studied by Neuman et al. [10]. After several passages in cell culture, virions escape the inhibition of replication but show attenuated forms. Rangan et al. [11], described highly structured areas of RNA that might be less accessible to complementary oligomers, but these convoluted areas would provide small binding sites for conventional drug molecules; therefore, a combination of scanning for structure and sequence conservation may be appropriate to find therapeu-
tic targets. Previous studies using in silico methods found drug-like molecules that would inhibit SARS-CoV replication by action on the frameshifting region at the overlap between ORF1a and ORF1b [12]. The same molecule has been shown to affect replication in SARS-CoV-2 [9].

In this work, we scanned the SARS-CoV-2 genome to seek for novel likely critical areas for virus replication focusing on frameshifting predictors. We explored the likely biological relevance of this feature through the study of sequence conservation and its suitability as a potential drug target by the analysis of the structural properties and the drug docking prediction.

MATERIAL AND METHODS

Genomewide frameshifting signal search. A prediction of the relevant sequence and structures in the viral reference genome for SARS-CoV-2 (NC 045512) [13] was performed using the KnotInFrame tool [14]. The output determined the sequence and position of slippery sequences and nearby pseudoknots, since both criteria are needed to predict frameshifting. Our focus on a particular region was established by combining KnotInFrame output with biological knowledge. We focused on previously undescribed frameshifting regions and the likely regulatory roles of UTR regions. Once a sequence of interest met these criteria, an inspection of the predicted secondary structure was achieved with additional tools ipknot [15] and RNAfold from Vienna Suite [16]. The secondary structure for the segment of interest in dot bracket notation was chosen from the inspection of the overall conformation of the 5’UTR and assuring to include the slippery region. The likelihood of the secondary structure was assessed by computing the minimum free energy (MFE) of a large number of random sequences of SARS-CoV-2 of the same length as the sequence of interest into mFold, in order to obtain an empirical distribution of MFE and so assess how dominant the proposed structure would be [17].

Conservation of the sequence of interest. Sequence conservation was assessed for the sequence of interest deter-
minded in the previous step as a reliable trait of biological relevance. The conservation of the sequence of interest was evaluated in two steps. First, the conservation between SARS-CoV-2 and other human and animal hosted coronavirus genomes was studied by the computation of a cladogram and by the search for the alignment of the sequence of interest against a comprehensive viral database. A total of 21 high quality genomes from coronavirus hosted in humans and other species were selected based on subjective criteria regarding variability and relevant facts to build a cladogram. The genomes were downloaded from GenBank and aligned with Clustal Omega [18] using the default parameters. The cladogram was constructed using a maximum likelihood estimate with FastTree [19], under a GJT model of nucleotide evolution. The package ggtree [20] was used in R [21] to generate the graphic of the cladogram and the multiple sequence alignment (MSA). In addition to this alignment of the SARS-CoV-2 and another 20 coronavirus genomes, the sequence of interest was examined by ViroBLAST [22]. This tool provides a blastn [23] alignment with a comprehensive database of all types of virus, so that we would assess any casual homology with any other virus. Secondly, we evaluated the conservation of the sequence of interest within SARS-CoV-2 isolates from different geographic locations since the onset of the pandemic. We took advantage of the fast contribution of genomes into the GISAID database. We filtered the genomes in the database in order to retain only high quality records (length greater than 29,000 nt and with a low number of undetermined positions). The number of variant site strains was assessed by blastn [23], making the distinction of variants at the whole 5'UTR region (1-265 nucleotide positions); and the number of variants at the position of the sequence of interest. Further individual inspections of mismatched genomes were conducted to ensure whether the variation was not due to technical sequencing reasons.

**Prediction of 3D structure and molecular docking.** Upon consideration of different alternatives, the structure of the sequence of interest in dot-bracket notation and the underlying nucleotide sequence were imported into Rnacomposer [24] to obtain a 3D structure prediction in .pdb format. The file in .pdb format was used as input for the virtual scan for active sites. This task was carried out using Autodock tools suite [25]. This suite comprises the AutoGrid and Autoligand tools for the search of active sites in a molecular 3D structure. A combination of manual selection of the region of interest and automatic search space by the tools was used to obtain the coordinates and dimensions of a putative active site. These data were used as inputs for the molecular docking by the Autodock Vina tool. The virtual docking tested the binding of a set of molecules specially selected for drug screening: the NCBI maximum diversity set II. The affinity of molecules to bind the active site was assessed by the minimum free energy in Kcal/mol.

### RESULTS

**Frameshifting prediction.** We used Knotinframe program [14] to detect cis-acting signals, the nucleotide sequence and the position of the heptameric slippery sites and near pseudoknots as prediction for −1 PRF. A list of genomic regions of SARS-CoV-2 (NC 045512.2) where a frameshifting signal was predicted by the KnotinFrame program is shown in Table 1. The stability of every predicted structure is also indicated by the MFE value, on the rightmost column. A more negative value of MFE represents a more stable and likely to be a functional structure. This value is mainly dependent on the length of the sequence. The pseudoknot we propose associated with the pattern sequence at position 76 (UUUAAAA) that was identified as −1 PRF, is ranked fourth and exhibit the lowest MFE value of −34.80.

<table>
<thead>
<tr>
<th>Slippy sequence</th>
<th>Slippery pos.</th>
<th>Pseudoknot start</th>
<th>Pseudoknot end</th>
<th>Length</th>
<th>Deltarel</th>
<th>MFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTTAAC</td>
<td>13462</td>
<td>13469</td>
<td>13549</td>
<td>80</td>
<td>0.126</td>
<td>-34.80</td>
</tr>
<tr>
<td>GGGTTTA</td>
<td>4261</td>
<td>4268</td>
<td>4328</td>
<td>60</td>
<td>0.092</td>
<td>-15.60</td>
</tr>
<tr>
<td>AAATTIG</td>
<td>6071</td>
<td>6078</td>
<td>6158</td>
<td>80</td>
<td>0.076</td>
<td>-16.10</td>
</tr>
<tr>
<td>TTTAAAA*</td>
<td>76</td>
<td>83</td>
<td>123</td>
<td>40</td>
<td>0.070</td>
<td>-14.00</td>
</tr>
<tr>
<td>GGGTTTIT</td>
<td>13348</td>
<td>13355</td>
<td>13475</td>
<td>120</td>
<td>0.051</td>
<td>-34.30</td>
</tr>
<tr>
<td>GGGTTITG</td>
<td>8183</td>
<td>8190</td>
<td>8270</td>
<td>80</td>
<td>0.049</td>
<td>-15.10</td>
</tr>
<tr>
<td>TTTAAT</td>
<td>4264</td>
<td>4271</td>
<td>4331</td>
<td>60</td>
<td>0.047</td>
<td>-15.60</td>
</tr>
<tr>
<td>CCCAAAAA</td>
<td>20646</td>
<td>20655</td>
<td>20773</td>
<td>120</td>
<td>0.046</td>
<td>-29.00</td>
</tr>
<tr>
<td>TTTAAAAA</td>
<td>6514</td>
<td>6521</td>
<td>6621</td>
<td>100</td>
<td>0.038</td>
<td>-19.40</td>
</tr>
<tr>
<td>TTTAAC</td>
<td>20817</td>
<td>20824</td>
<td>20924</td>
<td>100</td>
<td>0.035</td>
<td>-30.20</td>
</tr>
<tr>
<td>TTTTTT</td>
<td>11076</td>
<td>11085</td>
<td>11183</td>
<td>100</td>
<td>0.035</td>
<td>-19.60</td>
</tr>
</tbody>
</table>

*The line in bold face was chosen as the sequence of interest. MFE: minimum free energy.*

Table 1. Summary of the output of KnotinFrame tool on NC_045512.2 genome.
different programs in dot bracket notation. The IPknot tool [15] has been used in two fashions, firstly by the only input of the sequence of interest and secondly, by the input of the whole 5'UTR region, and then cutting out the prediction for positions 7:130. In both cases, IPknot predicts a knotted structure just downstream of the slippery region (the pattern of opening and closing brackets do not match, meaning that stem-loops bind to outer regions). Interestingly, the prediction we obtained using IPknot fully agrees with the prediction obtained recently in that genomic area using the Rosetta tool [11].

Similarly, the secondary structure of the whole 5'UTR regions was also obtained by IPknot tool and the graphical representation of that secondary structure using the VARNA software [26] is shown in Figure 2. Upon the inspection of the secondary structure, a sequence of interest spanning from position 74:130 was selected (left top corner) to include the slippery heptamer at position 76. The selected predicted frameshifting region clearly falls into the 5'UTR of NC_045512.2 reference genome [14], which spans from 1 to 265 nt as the first start codon for the coding sequence is at 266 position in SARS-CoV-2. However, if −1 PRF occurs within the 5'UTR region, probably at the U nucleotide at position 95 nt, then there is an upstream AUG codon at position 107 that can act as start codon and viral translation might be altered. The sequence of interest spanning from position 74:130 was selected and it is shown in Figure 1 along with the secondary structure as predicted by

![Figure 1](image1.png)

**Figure 1** Segment of 5'UTR NC 045512 sequence and different predicted secondary structures. The slippery heptamer is highlighted in yellow.

![Figure 2](image2.png)

**Figure 2** Secondary structure of 5'UTR region of NC 045512 as predicted by IPKnot. Region of interest framed on the top left. Slippery sequence in yellow.
Putative Secondary Structure at 5'UTR as a Potential Antiviral Target against SARS-CoV-2

E. Garcia-Moran, et al.

As indicated by its computed MFE, we analyzed 1,509 random sequences from NC_45512.2 of the same length as the sequence of interest. Their MFE values were computed by mFold [17] to obtain an empirical distribution. The predicted value for the sequence of interest (-11 Kcal/mol) was ranked among the 25% lowest values, indicating that it is likely to form a stable secondary structure.

The slippery sequence and neighboring structured segment. The sequence (positions 74—130) was also used to proceed with the analysis so that it includes the slippery region and the structure of the stem and loop and the pseudoknot.

In order to test the probability of the predicted secondary structure as indicated by its computed MFE, we analyzed 1,509 random sequences from NC_45512.2 of the same length as the sequence of interest. Their MFE values were computed by mFold [17] to obtain an empirical distribution. The predicted value for the sequence of interest (-11 Kcal/mol) was ranked among the 25% lowest values, indicating that it is likely to form a stable secondary structure.
The conservation of the 74--130 region among Coronaviridae family is shown in Figure 4. Interestingly, while this region was identical in all the isolates from SARS-CoV-2 including isolates from human patients from distant geographical localizations (MT370831, New York; and ...
LC542809, Japan) and from animals suspected to be infected from humans (MT396266, farm mink; and MT365033, zoo tiger), increasing differences were observed in other *Coronaviridae*. A minor difference in one nucleotide was found in a bat sequence (MT996532), while the differences increased in the pangolin hosted virus, and SARS-CoV-1 (NC\_004718, Tor2 strain). Subsequently, we also tested the sequence of interest for similarity to other viruses on the Viroblast database [22]. The search parameters were kept at nominal values except for the word length, which was changed from ten to seven to increase the likelihood of matching slippery sequences. Viroblast search yielded 10 hits derived from two sources (Table 2): the top four hits were coronaviruses isolated in China from different species of bat *Rhinolophus pusillus/sinicus* between 2011 and 2017 and the latest hits a mouse-adapted laboratory model derived from SARS-CoV (consecutive GenBank accession numbers from HQ890526 to HQ890531.1; although these isolates were sampled from the Urbani strain GenBank AY278741 rather than Tor2). Finally, we evaluated the conservation of the sequence of interest among clinical SARS-CoV-2 isolates. The 5’UTR (1-265 nucleotide positions) of NC\_045512.2 was searched using BLASTN [23] analysis against 54,466 high quality filtered genomes (out of 84,140) retrieved from GISAID on the 21st of August, 2020. While, only a 19.8% of the genomes (10,789 out of 54,466) had a 100% identity in the complete 5’UTR, the 74–130 region was highly conserved, as it was 100% identical in 99.3% of SARS-CoV-2 genomes tested (53,077 out of 53,456).

**RNA 3D structure.** In order to carry out molecular docking on potential active sites of the sequence of interest was continued the nucleotide sequence in the Figure 1 and the selected region of the secondary structure in Figure 2 as inputs into RNAcomposer [24] to obtain a .pdb file of the nucleotide sequence. The results of the predicted 3D structure and the possible location of a drug binding site are shown in Figure 5. The volume of the binding site is the result of the exploration with Autodock tools. The coordinates of the binding site were obtained in .pdb format and they were passed into Autodock Vina (the exhaustiveness search parameter at a default value of eight; and the random seed sequence was fixed) [25]. The results of the docking by Autodock Vina against the lead compounds from the NCI Maximum Diversity set II are shown in Table 3. The Autodock Vina predicted the affinity of the lead compounds by the computation of the MFE. As more negative values of MFE mean higher binding affinities, so the lead compounds were ranked by this value. The number of hydrogen bond donors and acceptors and the molecular weight in g/mol are annotation data from PubChem. These data show how likely a compound is to be used as an oral drug [27].

**DISCUSSION**

This study has revealed a previously unnoticed feature in the SARS-CoV-2 genome, which is likely to play a biological role on account of the remarkable conservation of its sequence and stability of the structure. The close occurrence of the slippery sequence and a likely stable pseudoknot suggests that this may be an area of frameshifting, in addition to the previously described overlapping region of ORF1a and ORF1b, where frameshifting has been proven for SARS-CoV [8] and also present in SARS-CoV-2 [9]. We focused on a different region, previously unnoticed in the 5’UTR. The fact that no protein may be linked with the sequence may argue against frameshifting, as may that of the overlap between ORF1a and ORF1b. Supporting the role of 5’UTR, Zhu et al. [28] demonstrated that different natural deletions in the 5’UTR of FMDV (foot-and-mouth disease virus) markedly affected the pathogenicity and species tropism of the virus. Frameshifting linked with 5’UTR has been described in HIV-1 [29], and in this case the structure next to the slippery sequence is a stem and loop, without additional pseudoknotting.

Another important endeavour of this work is to consider this RNA structured area as a useful target for feasible drug
intervention. Puzzlingly, the description of a possible drug against the pseudoknot involved in frameshifting between ORF1a and ORF1b in SARS-CoV did not progress to an actual drug for use in health care [12,30], probably due to the lag in time of this discovery after the 2003 SARS-CoV outbreak. The same molecule that was found to inhibit viral replication of SARS-CoV appears to be effective against SARS-CoV-2 [9].

Rangan et al. [11] performed a wide analysis on the SARS-CoV-2 and SARSr-CoV genomes, and they classified multiple regions in terms of the conservation and RNA structure. In agreement with our approach, they consider that structured regions would be ideal targets for small drug molecules. In Table 2 (eighth row) of their article, they describe, among others, sequence 40:157 of NC\textunderscore identifier:045512-2 as highly conserved and structured. We reproduce in Figure 1 their proposal of structure for our region of interest. The result of the alignment of our sequence of interest against the Viroblast database showed that the sequence may have been close to SARS-CoV as described in 2003 [31]. We found two pathways of highly similar sequences; coronaviruses isolated from bats in the following years and from laboratory-derived strains developed to create a mouse-adapted model from the Urbani strain of SARS-CoV-2. Although outside of this work, these findings support the role of bats as intermediate hosts between SARS-CoV and SARS-CoV-2, and the possibility that some unrecognized variation in strains of SARS-CoV would manifest relevant features as with the sequence which we describe.

A limitation of our work is that it was restricted to computational analysis. This shortcoming is likely more relevant when it comes to the determination of the tridimensional structure of RNA and its subsequent docking. The determination of the crystal structure 3D prediction and drug docking has been developed for proteins rather than RNA. One of the features of RNA that makes docking difficult is its flexibility. However, successful discovery of ligands against SARS-CoV-2 pseudoknot by a computed 3D structure has been described before [12,30]. Clinical evidence of pharmacological actions against RNA viral genomes was achieved by drugs such as sofosbuvir (tradename Sovaldi) against Hepatitis C Virus (HCV). These drugs are described as nucleotide analogues. They bind to the target region as complementary sequence would do but they differ from short chains of nucleotides so that they may resist lytic enzymes.

Our screening for drug ligands was an exploratory analysis, as it was limited to 1,507 compounds from NCBI maximum diversity set II. In Table 3 several compounds have a MFE lower than -10 Kcal/mol, though by a small margin. This fact does not preclude oral activity. NSC61610 was given orally, once a day to mice in an experimental model of H1N1 influenza infection [32]. The mice had less mortality and the response was better than with tamiflu after the sixth day of infection. However, that mechanism of action is unrelated to interactions with viral RNA, as NSC61610 acts as a modulator of the immune response.

In conclusion, we have identified a relevant sequence in the 5'UTR region of SARS-CoV-2. It displays traits which have a high potential for playing an important role, either through frameshifting or other mechanism. A remarkable conservation within SARS-CoV-2 isolates strongly supports a biological role for this sequence. Our analysis of the drug susceptibility of this sequence is hindered by the inconsistent predictions of bioinformatic tools. It is however very likely that a strong structure of this area will allow effective action of relatively simple drug molecules.

ACKNOWLEDGMENTS

We kindly acknowledge Dr. Nigel Cook for English revision.

FUNDING

None to declare.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES


