Description of a Novel Frameshifting Site in the 5'UTR of SARS-CoV-2 as a Potential Drug Target

Emilio García-Moran¹,*, Marta Hernandez², David Abad², Jose Mª Eiros³

1 Independent scholar.
2 Laboratorio de Biología Molecular y Microbiología, Instituto Tecnológico Agrario de Castilla y León, 47071 Valladolid, Spain.
3 Centro Nacional de la Gripe. Facultad de Medicina, Microbiology Department. 47005 Valladolid, Spain.

corresponding author: *egarmo@egarmo.com

Abstract

SARS-CoV-2 is an enveloped positive-sense single-stranded RNA coronavirus that causes COVID-19 whose present outbreak has cost a high number of casualties throughout the world. The aim of this work was to scan the SARS-CoV-2 genome in search for new 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 frameshifting. The potential biological value of this region is shown by its similarity with other coronaviruses related with SARS-CoV and its sequence conservation within isolates from SARS-CoV-2. We have predicted the secondary structure of the region by means of different bioinformatic tools. We have chosen a probable secondary structure to proceed with a 3D reconstruction of the structured segment. We carried out virtual dockin on the 3D structure to look for a binding site and then for drug ligands from a database of lead compounds. Several molecules that would probably administered as oral drugs show promising binding affinity within the structured region and so it would be possible interfere the potential regulatory role of our sequence of interest.

Keywords

SARS-CoV-2; frameshifting; drug screening; bioinformatics

Declarations

Funding: not applicable.
Conflict of interest/competing interest: Emilio García Morán, Marta Hernández, David Abad and José Mª Eiros declare that they have no conflict of interest.
Availability of data and code: please contact the author for data requests.
Author’s contributions. Emilio García Morán conceived and designed the study, wrote the paper and performed research. Marta Hernandez wrote the paper and performed research. David Abad performed research and analyzed data. José Mª Eiros performed research.
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 mid September 2020, the pandemic outbreak has caused nearly a million deaths worldwide.

SARS-CoV-2 belongs to the Coronaviridae family and is related to SARS-CoV and Middle East respiratory syndrome (MERS)-CoV (79% and 50% genomic similarity, respectively). SARS-CoV caused an epidemic outbreak in 2003 and MERS caused an outbreak in 2012 [4]. All three belong to the Betacoronavirus genus. Coronaviruses cause zoonotic infections, so they may pass from a host species to a different one through small changes in their genome. SARS-CoV-2 demonstrated a high genetic similarity (more than 85%) to a group of virus known as SARS related coronavirus (SARSr-CoV), which are isolated from animal hosts, including Hipposideros bats and pangolins (Manis javanica). These species seems to be candidates of being intermediate hosts for SARS-CoV-2 [29] [9].

These viruses have a positively translated single strand RNA genome and they use programmed −1 ribosomal frameshifting (−1 PRF) to direct the synthesis of immediate early proteins that prepare the infected cell for takeover by the virus. Frameshifting is a smart mechanism for the translation of a genomic sequence into two different proteins by moving the translation frame one position (backwards or forward) in the union between RNA and the ribosome [10]. A typical frameshifting signal has two essential elements: a characteristic 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 [2]. 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 been developed to predict these structures [13]. The mechanism of action of pseudoknots is not completely understood; some authors suggested 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. [19] 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. [11] 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 [11]. Starting with early studies, point mutations at the slippery sequence have proved to have an important effect on viral replication [19]; thus, they would 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. [16]. After several passes of viral culture, virions escape the inhibition of replication but show attenuated forms. Rangan et al. [23], 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 therapeutic 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 [17]. The same molecule has been shown to affect replication in SARS-CoV-2 [11].

In this work, we scanned the SARS-CoV2 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.

Materials and Methods

0.1 Genomewide frameshifting signal search

A prediction of the relevant sequence and structures in the viral reference genome for SARS-CoV-2 (NC_045512) [30] was performed using the KnotInFrame tool [27]. 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 [25] and RNAfold from Vienna Suite [15]. 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. [8].

0.2 Conservation of the sequence of interest

Sequence conservation was assessed for the sequence of interest determined 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 [26] using the default parameters. The cladogram was constructed using a maximum likelihood estimate with FastTree [21], under a GTT model of nucleotide evolution. The package ggtree [31] was used in R [22] 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[6]. This tool provides a blastn 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[1], 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.

0.3 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 rnamcomposer[20] 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[28]. 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

A list of genome regions of SARS-CoV-2 (NC_045512.2) where a frameshifting signal was predicted by the KnotInFrame program is shown in Table 1.

| slippery sequence | slippery pos. | pseudoknot start | pseudoknot end | length | Deltarad | MFE     |
|------------------|---------------|------------------|----------------|--------|----------|---------|
| TTTAAC           | 13462         | 13469            | 13549          | 80     | 0.126    | -34.80  |
| GGGTTTA          | 4261          | 4268             | 4328           | 60     | 0.092    | -15.60  |
| AAATTGG          | 6071          | 6078             | 6158           | 80     | 0.076    | -16.10  |
| TTTAAAA          | **76**        | **83**           | **123**        | **40** | **0.070**| **-14.00**|
| GGGTTTT          | 13348         | 13355            | 13475          | 120    | 0.051    | -34.90  |
| GGGTTTG          | 8183          | 8190             | 8270           | 80     | 0.049    | -15.10  |
| TTTAAAT          | 4264          | 4271             | 4331           | 60     | 0.047    | -15.60  |
| CCCAAAAA         | 20646         | 20655            | 20773          | 120    | 0.046    | -29.00  |
| TTTAAAAA         | 6514          | 6521             | 6621           | 100    | 0.038    | -19.40  |
| TTTAAAC          | 20817         | 20824            | 20924          | 100    | 0.035    | -30.20  |
| TTTTTTT          | 11076         | 11085            | 11183          | 100    | 0.035    | -19.60  |

Table 1. Summary of the output of the KnotInFrame tool on the NC_045512.2 genome. The line in bold face was chosen as the sequence of interest.

Knotinframe provides the nucleotides and position of the slippery sequence and the near pseudoknot. The predicted stability of every predicted structure is indicated by the MFE value, on the rightmost column. A more negative value of MFE represents a more stable and likely structure. This value is mainly dependent on the length of the
sequence. The pseudoknot associated with the pattern sequence (TTTAAAA) at position 76 is ranked fourth in terms of MFE among the predictions on the whole genome. However, all other structured sequences are longer and this causes their stability to not be so much significantly higher than the region associated with the slippery sequence at position 76.

The frameshifting region clearly falls within the 5’UTR of NC_045512.2 reference genome, which spans from 1 to 266 nt. The nucleotide sequence of the sequence of interest is shown in Figure 1 along with the secondary structure as predicted by different bioinformatic tools. There is a clear disagreement between these predictions and not all of them actually predict a pseudoknot. The IPknot tools was used in two fashions, one by the input of the sequence of interest only and second, 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). We include a published prediction of the area in the first line. The prediction by knotinframe falls back on tools from the Vienna suite, namely RNAfold and there is a clear disagreement, since RNAfold This disagreement may be due to the fact that Knotinframetool inputs a shorter sequence, so, all the outside bindings predicted between positions 90-100 may change when a longer sequence is considered and further downstream complementary segments come into play.

Figure 1. Segment of 5’UTR NC_045512 sequence and different predicted secondary structures. The slippery heptamer is highlighted in yellow.

![Secondary structures](image)

The secondary structure of the whole 5’UTR regions was obtained by two different tools: ipKnot Figure 3 and RNAfold Figure 2. In both cases the segment starting at the position of the slippery sequence (bp 75 onwards) is placed at the top left. There is a clear disagreement between the prediction of both figures, a knotted structure is predicted only by IPKnot.

Upon the inspection of the secondary structure shown in Figures 3 and 2 a sequence of interest spanning from position 74:130 was selected to include the slippery sequence and neighbouring structured segment. The secondary structure was cut off from the prediction by IPknot. It is shown in detail in figure 4.

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 to obtain an empirical distribution. The predicted value for the sequence of interest (-11 Kcal/mol) was ranked within empirical distribution of the MFE values. The histogram and frequency curve of this distribution is shown in Figure 5. The vertical line is set at -11 Kcal/mol. Clearly, few random sequences show this value. This value was at the top 5% of the negative endo of the distribution. This sign that the predicted structure is fairly stable in relation with other segments of NC_45512 supports that this sequence may occur in the predicted form of a stem-loop with outer bindings to form a pseudoknot and thus, along with the immediate slippery sequence form a frameshifting signal.
Figure 2. Secondary structure of 5'UTR region of NC_045512 as predicted by RNAfold. Region of interest framed on the top left. Slippery sequence in yellow.

The sequence spanning from position 74–130 was used to proceed with the analysis so that it would include the slippery region and the structure of the stem and loop and the pseudoknot. The graphical representation of that secondary structure using the VARNA software [5] is shown in Figure 4.

Conservation of the sequence of interest in relation to a set of other coronaviruses is shown in Figure 6. A multiple sequence alignment of the 5'UTR is shown on the right side, with a zoomed in view on the coordinates around the sequence of interest. All the isolates from SARS-CoV-2 were identical as far as the sequence of interest is concerned. This includes the samples from human patients in distant places (MT370831, New York; and LC542809, Japan) and isolates from animals that were suspected of being infected from humans (MT396266, farm mink; and MT365033, zoo tiger). A minor difference in one nucleotide was found in a bat sequence (MT996532) and then the differences increased with the pangolin hosted virus, and SARS-CoV from the 2003 outbreak (NC_004718, Tor2 strain). The differences are shown for MERS and other coronaviruses.

The input of the sequence of interest on the Viroblast database yielded 10 hits. 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. The ten hits may be summarized in two sources. A group of sequences with consecutive GenBank accession numbers from HQ890526 to HQ890531.1 came from a mouse-adapted laboratory model derived from SARS-CoV, although this was sampled from the Urbani strain (GenBank AY278741) rather than Tor2. The top four hits were isolated in China from different species of bat (Rhinolophus pusillus/sinicus) between 2011 and 2017.

The lower panel of the figure shows the predicted secondary structure of the whole area where frameshifting is predicted. The links outside the more distal stem-loop represent the actual pseudoknot. The slippery region is highlighted in green at position 76.

We evaluated the conservation of the sequence of interest between clinical isolates of SARS-CoV-2. The 5’UTR (1-265 nucleotide positions) of NC_045512.2 was searched using BLASTN analysis against 54,466 high quality filtered genomes (out of 84,140) retrieved from GISAID on the 21st of August, 2020. We found that 10,789 out of 54,466 (19.80%) genomes had a 100% identity, whereas the BLASTN analysis restricted to the
sequence of interest (NC_045512:2.70-123) showed that this region was highly conserved, as it was 100% identical in 53,077/53,456 (99.29%) of SARS-CoV-2 genomes.

In order to carry out molecular docking on possible active sites of the sequence of interest was continued the nucleotide sequence (Figure 1) and secondary structure (Figure 4) as inputs into RNAcomposer [20] 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 7. 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). The results of the docking by Autodock Vina against the lead compounds from the NCBI Maximum Diversity set II are shown in Table 3. The values in the first and second columns are the identifiers that link identification in the NCBI database (NSC code) with a full annotation in PubChem (https://pubchem.ncbi.nlm.nih.gov/).

The Autodock Vina predicted the affinity of the lead compounds by the computation of the MFE shown in the third column for the best matching lead

Table 2. List of similar hits to NC_045512:74-130 in Viroblast database.
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 are show how likely a compound is to be used as an oral drug. The was consistent with an orally feasible drug.

Figures 7, 8 are representations of the same 3D reconstruction of the whole sequence of interest loading the .pdb file obtained from RNA composer tool into Pymol. Figure 7 shows an oblique 5’ to 3’ view and the outline of the nucleotides in an improved way. Distal to the slippery sequence lies the site of active binding as predicted by the AutoGrid tool. The area is shown as a rounded volume in blue.

Figure 8 is a side view of the surface representation of the same structure. The slippery region is coded in red. The active site holds a bound drug ligand (NSC308835/pubChem328761, Table 3 second line) as predicted in the docking analysis. This compound is highlighted on account of its top second affinity with the active site and the compliance of every Lipinski’s rule of five to be an orally active drug.

The results of the docking performed by Autodock Vina between the active site...
Figure 5. The histogram and density curve of the minimum free energy values of random sequences from NC_45512.2 of the same length as the sequence of interest. The vertical line at -11 Kcal/mol shows the computed MFE for the sequence of interest.

shown in Figure 7 and the drug compound set NCBI maximum diversity set II are summarized in Table 8.
Figure 6. A cladogram and multiple sequence alignment. The middle part shows the alignment of the region in 21 coronaviruses. On the left, are shown the groups of sequences in terms of similarity. A zoomed in view on the region of the sequence of interest is shown on the right.

| NSC id  | pubChem id | MFE   | Molecular formula | H bond donors | H bond acceptors | active torsions | Mol weight |
|---------|------------|-------|-------------------|---------------|------------------|----------------|------------|
| 293778  | 325266     | -12.2 | C_{40}H_{26}N_4S  | 0             | 5                | 6              | 594.7      |
| 308835  | 328761     | -11.1 | C_{30}H_{32}N_2O_4 | 0             | 4                | 3              | 484.6      |
| 61610   | 247228     | -11.1 | C_{34}H_{24}N_4O_2 | 4             | 4                | 6              | 548.6      |
| 37641   | 235856     | -11   | C_{29}H_{33}FO_6  | 2             | 7                | 3              | 496.6      |
| 319990  | 330740     | -11.7 | C_{23}H_{33}N_2S_2 | 4             | 6                | 4              | 474.6      |
| 93354   | 261360     | -10.6 | C_{28}H_{22}NO_2S  | 1             | 4                | 2              | 447.6      |
| 122819  | 452548     | -10.5 | C_{32}H_{32}O_{23}S | 3             | 14               | 6              | 656.7      |
| 37553   | 235811     | -10.5 | C_{30}H_{28}N_4O_2 | 2             | 2                | 2              | 476.6      |
| 37641-b | 235856     | -10.5 | C_{29}H_{33}FO_6  | 2             | 7                | 3              | 496.6      |
| 80997-a | 255437     | -10.4 | C_{30}H_{36}N_2O_3 | 2             | 4                | 2              | 473        |

Table 3. Results of docking of lead compounds from NCI diversity set II against the predicted active site in the sequence of interest.
Figure 7. Graphical representation of the nucleotide backbone of the sequence of interest. The sequence of interest is shown from the 5’ end (left) to the 3’ end (right). The rounded purple volume in the middle shows the active site as predicted by the AutoDock suite tools. The slippery sequence is on the left bottom, in purple colour, the rest of nucleotide pieces are coded according to chemical composition.
Figure 8. Side view of the sequence of interest in a surface representation. The red area on the left shows the slippery sequence. The active ligand site holds one of the best matches: NSC308835/pubChem328761 (see Table [3]) in its docked position.
Discussion

This work shows a previously unnoticed feature in the SARS-CoV-2 genome, which is likely to play an important biological role on account of the remarkable conservation of its sequence. The close occurrence of the slippery sequence and likely a stable pseudoknot suggests that this may be an area of frameshifting, in addition to the overlapping region of ORF1a and ORF1b, where frameshifting has been proven for SARS-CoV \[19\] and also present in SARS-CoV-2 \[11\]. We have focused on a different region, previously unnoticed in 5’UTR. The fact that no protein may be linked with the sequence may argue against frameshifting as that of the overlap between ORF1a and ORF1b. Supporting the role of 5’UTR, Zhou et al. \[32\] 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 \[8\], 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 to use in health care \[18\] \[17\], 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-CoV2. \[11\]

Rangan et al. \[23\] 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\_045512-2 as highly conserved and structured. We reproduce in [1] their proposal of structure for our region of interest.

There is a clear disagreement between the secondary structures not only due to the use of different tools (Rosetta \[23\], RNAfold and IPKnot, but also depending whether the input to these tools was a large, wide scope region as the whole 5’ UTR or just a fragment. The observation of the differences between [3] and [2] shows that disagreement exists even for the overall 5’UTR structure. We decided to work on the local detach of the segment of interest as predicted by IPKnot.

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 \[24\]. We found two pathways of highly similar sequences, one in wildly occurring in isolation from bats in the following years and another one in laboratory derived strains to create a mouse adapted model from 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 hold relevant features as the sequence that we describe.

The primary limitation of our work is that it was restricted to in silico research. 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 \[18\] \[17\]. 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 a 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 1507 compounds from NCBI maximum diversity set II. In table 3, several compounds have a MFE lower than -10 Kcal/mol. This suggests that a search against a larger catalog would yield multiple candidates. As pointed out in the limitations, we fail to provide further evidence of the effect of a drug binding the pseudoknot as Park et al. did on SARS-Cov [18] or recently on SARS-CoV-2 [11]. Several compounds on the table would meet the criteria to be orally useful drugs according to Lipinski’s rule of five [14]. We point out the compound ranked second in terms of MFE affinity: NSC 308835 because it meets every Lipinski’s criteria. The next best compound did not meet that molecular weight should be less than 500 g/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 [12]. The mice had less mortality and the response was better than to 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.

As a final conclusion of this work, we claim to have found a relevant sequence in the 5’UTR region of SARS-CoV-2. It shows fairly temptative traits to play an important role, either through frameshifting or other mechanism. A remarkable conservation within the isolates in the present SARS-CoV-2 strongly support a biological role for this sequence. Our analysis of the druggability of this sequence is flawed by the inconsistent predictions of bioinformatic tools. It is however very likely that a strong structure of this area allows effective action of relatively simple drug molecules.

References

1. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215:403–410, Oct. 1990.

2. I. Brierley, S. Pennell, and R. J. C. Gilbert. Viral rna pseudoknots: versatile motifs in gene expression and replication. *Nature reviews. Microbiology*, 5:598–610, Aug. 2007.

3. J. Charbonneau, K. Gendron, G. Ferbeyre, and L. Brakier-Gingras. The 5’ utr of hiv-1 full-length mrna and the tat viral protein modulate the programmed -1 ribosomal frameshift that generates hiv-1 enzymes. *RNA (New York, N.Y.)*, 18:519–529, Mar. 2012.

4. C. M. Coleman and M. B. Frieman. Emergence of the middle east respiratory syndrome coronavirus. *PLoS pathogens*, 9:e1003595, 2013.

5. K. Darty, A. Denise, and Y. Ponty. Varna: Interactive drawing and editing of the rna secondary structure. *Bioinformatics (Oxford, England)*, 25:1974–1975, Aug. 2009.

6. W. Deng, D. C. Nickle, G. H. Learn, B. Maust, and J. I. Mullins. Viroblast: a stand-alone blast web server for flexible queries of multiple databases and user’s datasets. *Bioinformatics (Oxford, England)*, 23:2334–2336, Sept. 2007.

7. A. R. Fehr and S. Perlman. Coronaviruses: an overview of their replication and pathogenesis. *Methods in molecular biology (Clifton, N.J.)*, 1282:1–23, 2015.

8. A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neub?ck, and I. L. Hofacker. The Vienna RNA websuite. *Nucleic Acids Res.*, 36(Web Server issue):W70–74, Jul 2008.
9. G. Z. Han. Pangolins Harbor SARS-CoV-2-Related Coronaviruses. *Trends Microbiol.*, 28(7):515–517, 07 2020.

10. T. Jacks, H. D. Madhani, F. R. Masiarz, and H. E. Varmus. Signals for ribosomal frameshifting in the rous sarcoma virus gag-pol region. *Cell*, 55:447–458, Nov. 1988.

11. J. A. Kelly, A. N. Olson, K. Neupane, S. Munshi, J. San Emeterio, L. Pollack, M. T. Woodside, and J. D. Dinman. Structural and functional conservation of the programmed -1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *J. Biol. Chem.*, Jun 2020.

12. A. Leber, J. Bassaganya-Riera, N. Tubau-Juni, V. Zoccoli-Rodriguez, P. Lu, V. Godfrey, S. Kale, and R. Hontecillas. Lanthionine synthetase c-like 2 modulates immune responses to influenza virus infection. *Frontiers in immunology*, 8:178, 2017.

13. C. S. Lim and C. M. Brown. Know your enemy: Successful bioinformatic approaches to predict functional rna structures in viral rnas. *Frontiers in microbiology*, 8:2582, 2017.

14. C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 46:3–26, Mar. 2001.

15. R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. Vennarna package 2.0. *Algorithms for molecular biology : AMB*, 6:26, Nov. 2011.

16. B. W. Neuman, D. A. Stein, A. D. Kroeker, M. J. Churchill, A. M. Kim, P. Kuhn, P. Dawson, H. M. Moulton, R. K. Bestwick, P. L. Iversen, and M. J. Buchmeier. Inhibition, escape, and attenuated growth of severe acute respiratory syndrome coronavirus treated with antisense morpholino oligomers. *J. Virol.*, 79(15):9665–9676, Aug 2005.

17. H.-J. Park and S.-J. Park. Virtual Screening for RNA-Interacting Small Molecules, pages 235–252. 08 2012.

18. S. J. Park, Y. G. Kim, and H. J. Park. Identification of RNA pseudoknot-binding ligand that inhibits the -1 ribosomal frameshifting of SARS-coronavirus by structure-based virtual screening. *J. Am. Chem. Soc.*, 133(26):10094–10100, Jul 2011.

19. E. P. Plant, G. C. Pérez-Alvarado, J. L. Jacobs, B. Mukhopadhyay, M. Hennig, and J. D. Dinman. A three-stemmed mrna pseudoknot in the sars coronavirus frameshift signal. *PLOS Biology*, 3(6), 05 2005.

20. M. Popenda, M. Szachniuk, M. Antczak, K. J. Purzycka, P. Lukasiak, N. Bartol, J. Blazewicz, and R. W. Adamiak. Automated 3D structure composition for large RNAs. *Nucleic Acids Res.*, 40(14):e112, Aug 2012.

21. M. N. Price, P. S. Dehal, and A. P. Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26:1641–1650, July 2009.

22. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.
23. R. Rangan, I. N. Zheludev, and R. Das. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *RNA*, May 2020.

24. A. Roberts, D. Deming, C. D. Paddock, A. Cheng, B. Yount, L. Vogel, B. D. Herman, T. Sheahan, M. Heise, G. L. Genrich, S. R. Zaki, R. Baric, and K. Subbarao. A mouse-adapted sars-coronavirus causes disease and mortality in balb/c mice. *PLoS pathogens*, 3:e5, Jan. 2007.

25. K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):85–93, Jul 2011.

26. F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7:539, Oct. 2011.

27. C. Theis, J. Reeder, and R. Giegerich. Knotinframe: prediction of -1 ribosomal frameshift events. *Nucleic acids research*, 36:6013–6020, Oct. 2008.

28. O. Trott and A. J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31(2):455–461, Jan 2010.

29. H. Wang, X. Li, T. Li, S. Zhang, L. Wang, X. Wu, and J. Liu. The genetic sequence, origin, and diagnosis of SARS-CoV-2. *Eur. J. Clin. Microbiol. Infect. Dis.*, Apr 2020.

30. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, and Y.-Z. Zhang. A new coronavirus associated with human respiratory disease in china. *Nature*, 579:265–269, Mar. 2020.

31. G. Yu. Using ggtree to visualize data on tree-like structures. *Current protocols in bioinformatics*, 69:e96, Mar. 2020.

32. Z. Zhu, F. Yang, W. Cao, H. Liu, K. Zhang, H. Tian, W. Dang, J. He, J. Guo, X. Liu, and H. Zheng. The pseudoknot region of the 5' untranslated region is a determinant of viral tropism and virulence of foot-and-mouth disease virus. *Journal of Virology*, 93(8), 2019.