Occurrence of the Slippery Sequence UUUAAAC in the RNA Genome that Generates the ORF1ab Protein of SARS-CoV-2

Geurdes H*

GDS Applied Mathematics BV, Den Haag, Netherlands

ABSTRACT

In the present brief report we look into the slippery sequence TTTAAAC (in cDNA format) of the ORF1ab protein of SARS-CoV-2. We found a number of TTTAAAC sequences where only one is actively producing a shift-1. There are three other sequences exactly positioned in the read-through of mRNA as the aforementioned. They do not produce a-1 frameshift. There is one position where in addition a pseudoknot occurs but no frameshift. We ask if it is possible to enforce or prevent shifts in TTTAAAC to destroy the ORF1ab derived proteins such as RNA-dependent RNA polymerase and/or 2'-O-ribose methyltransferase. Finally an mRNA polymer repressor of the one single effective frameshift is proposed for further research into a medicinal treatment. Perhaps that there are specific protein repressors.

Keywords: SARS-CoV-2; RNA Genome; mRNA; Frameshift

INTRODUCTION AND PROBLEM

The virus SARS-CoV-2 belongs to the family of Corona viruses from the order of the Nidovirales [1]. The virus SARS-CoV-2 causes the illness COVID19. A cure for COVID19 could be to find ways to suppress the propagation of SARS-CoV-2 infection. Here we look into a possible vulnerability of frameshift-1 in the ORF1ab protein that embraces e.g. the infection necessary RNA dependent RNA polymerase enzyme.

In Corona viruses the RNA sequence UUUAAAC is identified as the slippery sequence that enables changes in the genetic read-through [2]. It is known that a negative RNA is formed as a template to create new viruses [3]. This frame shift could be a vulnerable step in the biosynthesis of de novo virus and so a careful look at the RNA genetics seems to be in order. In the cDNA representation (GenBank MT419837.1) at hand we will look at TTTAAAC. This identification will be used when convenient.

FRAMESHIFT IN CORONA VIRUS

In the infection of SARS-CoV-2 in the host cell, extensive use is made of ORF1ab derived proteins. We mention e.g. the synthesis of new mRNA with the help of the ORF1ab derived enzyme RNA-dependent RNA polymerase.

With the use of a computer program we were able to simulate the synthesis of the ORF1ab from the genetic code in the SARS-CoV-2 data of GenBank: MT419837.1. In the synthesis, one shift-1 slippery sequence is apparently active for amino acid residue 4402 (genetic identifier 13462).

This TTTAAAC could represent a vulnerability of ORF1ab because if, in the computer model the shift is ignored, a totally different protein arises from the genetics. In addition, we found rubbish genetic code further downstream. So if there is no shift-1, then the synthesis of ORF1ab products will be destroyed further down to the 3' end starting from the protein residue 4402. Moreover, if there are similar inactive TTTAAAC sequences then perhaps there are other vulnerabilities in the ORF1ab synthesis. These vulnerabilities will also generate rubbish genetic codes and deactivate the ORF1ab. In the present report we raise a number of questions about the TTTAAAC sequences in the synthesis of ORF1ab.

For a-1 frameshift, a not too far away downstream pseudoknot in the RNA is necessary [2]. A pseudoknot can be defined as [4] two helical structures connected by two single-stranded loops. There can however also be other architectures as well [4,5]. Typically, the pseudoknot architecture is 6-9 nucleotides

separated from the signal code, which in the present case is (in cDNA format) TTTAAAC.

RESULTS

We observed the TTTAAAC in a number of locations, such as 1664, 6085, 6745 and 13462. Only for the slippery sequence TTTAAAC on 13462 there is a-1 shift in the read-through of the mRNA (in cDNA representation of GenBank: MT419837.1). The following points can be raised.

Why only at 13462

Note that only the 1664 ...TTTAAAC... generates the FK... in protein residues in the ORF1ab mRNA (cDNA rep).

So, if the slippery sequence is exactly in the protein generating read-through of 3 nucleotides of mRNA, we must have FKL, FKP, FKH, FKQ and FKR in the protein residues.

For ORF1ab this "in the protein generating read-through" slippery sequence is apparently ineffective for a shift-1.

The shift-1 such as in 13462 has the TTTAAAC starting at the third TTT of Phe=F. So, for ORF1ab, this could be an effective starting point for a slippery sequence with a shift-1. In terms of protein, the residues are: F, L, S, C, P, H, R, I, V, A and G that can start the sequence. The second one is then always TTA=Leu=L and the third one is then always AAC=Asn=N.

In the Table below the slippery sequences and the genetic sequence plus position in the read-through together with a small sequence of protein residues, are presented where interesting. It is remarkable that other TTTAAAC in precisely the same read-through position as 13462, produce rubbish genetics further downstream. Some of the absence appears to be related to the absence of a pseudoknot structure.

After the shift-1 there is a different starting point for the consecutive TTTAAAC readouts.

The blue and red C of sequence position 13462 (it is just one C) indicates the shift-1 to explain the R in FLNRVC in the Protseq column. Above double line break pre-shift and below double line break post-shift.

mRNA location	Genetic sequence	Prot1	Prot2	Protseq	Shift	Pseudonot	Shift-1 protein
1664	TTTAAACTAATGAAGAG	TTT	AAA	FKLNEE	0	0	~
6085	GATTTAAACCAGTTAACT	GAT	TTA	DLNQLT	0	0	@ code
6745	TGTTTAAACCGTGTTTGT	TGT	TTA	CLNRVC	0	0	TVFKPCLY@
13462	TTTTTAAACGGGTTTGCG	TTT	TTA	FLNRVC	-1	1	RVC
16669	ACATTTAAACTGTCTTATG	~	~	~	0	~	~
18475	CAATTTAAACACCTCATACC	~	~	~	0	1	~
20227	GAATTTAAACCCAGGAGTC	~	~	~	0	~	~
20817	TATTTAAACACATTAACAT	~	~	~	0	~	SIFKHINISCT@

The @ indicates "untranslatable nucleotide triplet code", i.e. genetic rubbish.

In the genetics beyond 13462 the frameshift-1 must be included in the consideration of the start of the TTTAAAC and of the position of a possible pseudoknot structure. The search for TTTAAAC ignored this feature. The @ indicates "untranslatable nucleotide triplet code", i.e., genetic rubbish.

Table 1: Slippery sequence TTTAAAC in the mRNA of ORF1ab with and without shift-1 effect.

Looking at Table 1 the question remains why in the read-through of ORF1ab mRNA (cDNA) we only have one shift-1 for slippery sequence at location 13462 while at e.g. 6745 and 6085 and at 20817 this shift-1 does not occur. This is despite the fact that the LN[C..]nt part is present in those cases. Arguments like Gibbs energy [2] appear to be invalid because the TTTAAAC appears at the same position in effective and ineffective slippery sequences of ORF1ab. The pseudoknot is perhaps an explanatory ground but apparently not always.

In location 6085 immediately after the shift-1, nonsense genetic code arises. It can be imagined that immediate subsequent genetic rubbish will be somehow avoided. The question is how this is done. The absence of a proper pseudoknot might give more insight.

In location 6745 we have 8 normal genetic codes before rubbish is encountered. The affected protein is nsp3.

In location 20817 we have 11 normal genetic triplets before rubbish genetics occurs.

Are there ways to activate the non-active slippery sequences? If we do, a complete different ORF1ab set of proteins will be synthesized such that the nsp proteins and e.g. RNA-dependent RNA polymerase will not be synthesized or be active.

Are there ways to prevent the shift-1 in the 13462 slippery sequence?. We will go into that point later.

In the glycoprotein of GenBank: MT419837.1, there is a TTTAAAC sequence at 24436. This sequence has the same read-through occurrence as the shift-1 in ORF1ab. It is GCTTTAAACACGCTTGTTA with TTTAAAC on position 24436. But like e.g. the ORF1ab GATTTAAACCAGTTAACT of position 6085, there does not occur a frameshift-1 in the synthesis of the glycoprotein S spike.

Note in addition that for position 18475 a stem-loop-stem-loop RNA pseudoknot occurs. If we number the first T of TTTAAAC in 18475 as 1, then after C7 we have A8-C9-C10-T11 and then C12- A13-T14-A15 that is stem knotted with G24-T23-A22-T21 which is connected by the loop C16-C17-A18-C19-T20. Then the loop, continuing from G24, T25-A26- C27-A28-A29 with the second stem A30- G31-G32-A33, loop connected via C34-T35, to T39-C38- C37-T36 and it continues via G4. So we have here within range of 4 nucleotides an RNA pseudoknot and a slippery sequence but no -1 frameshift. Let us contrast this with the frameshift-1 TTTAAAC of 13462.

If we enumerate the first T of 13462 as 1 then after C7 the 9 length loop G8-G9-G10-T11-T12-T13-G14-C15-G16, then (first stem) G17-T18- G19-T20, connected by the loop A21-A22-G23-T24 to the second stem G25-C26-A27-G28 via C29-C30 loop to the second stem C31-G32-T33-C34. The combination is (25,31), (26,32), (27,33), (28,34). Then it goes via the loop T35-T36 to the (first stem) A37-C38-A39-C40 and it continues. The combination is (20,37), (19,38), (18,39), (40,17). It is interesting to note that the active pseudoknot for 13462 perhaps holds vulnerability further downstream.

We may conclude that the pseudoknot in 13462 is effective for -1 frameshift but has a weak spot in stem formation. The one in 18475 is not but the latter has almost all qualifications to be effective. This is remarkable although the relatively close vicinity of 18475 to 13462 could be an explanation for the suppression of a-1 frameshift in 18475. Note that after the immediate next TTTAAAC of the one in 13462, i.e. the one in 16669, there is no substantial RNA pseudoknot structure.

In this brief report we looked at TTTAAAC structures in the read-through of mRNA/cDNA such as the one in 13462. This is apparently the only case with an effective -1 frameshift. If the shift-1 is introduced in all those other cases, the code @=rubbish genetics, occurs after a number of valid nucleotide triplets. It is unlikely that there is an overseer process that reads ahead of protein synthesis. Moreover, the pseudoknots are there to create a chemical condition to locally provoke the -1 frameshift. However, if the @ occurs immediately after a shift-1, then one

can imagine that the slippery nucleotide sequence will be chemically avoided.

However, note also the presence of the proper downstream geometry [2,4,5] without the frameshift for position 18475. That leaves the question why sequences where only after 8 or 11 translated residues downstream, the rubbish code occurs, are ineffective.

With this result the hypothesis that the frameshift is triggered by an incomplete translocation of two nucleotides instead of three, due to the resistance of the upper stem of the frameshift stimulatory signal to unwinding, can be questioned. There is a downstream stem that could hamper the three nucleotides readout. Another point of view could be that specialized Ribosomes [6,7] are (co-) responsible for suppressing the -1 frameshift signaled in 18475 and allowing the one in 13462. This could also be an approach to change the -1 frameshift synthesis of ORF1ab proteins in SARS-CoV-2, especially when considering the G46-G47-C48-A49 *vs* C40-C41-G42-T43 possibility of stem formation and this stem appears too far away from the 13462 TTTAAAC.

In the above, we also go the other way around and ask if it is possible to prevent the 13426 shift-1. This will, like activating inactive TTTAAAC, a route to mess up the ORF1ab protein and the vital derived proteins thereof such as RNA-dependent RNA polymerase. It appears that a medicine against the in-host propagation of SARS-CoV-2 could be to interfere with its vulnerabilities of active and not active -1 ribosomal frameshifting in the read-through of the positive strand RNA.

Let us look at the deactivation of 13462 frameshifting in the read-through. Interestingly, in the domain of HIV research, scientists already wondered if there are cellular conditions that may modulate I guess [6]. Here we will look at the possibility of an mRNA type repressor.

A part of the RNA containing the 13462 sequence is (in RNA code)

UCGUUUUUAAACGGGUUUGCGGUG. This corresponds to the peptide FLNRVC=Phe-Leu-Asn-Arg-Val-Cys. This peptide sequence is a part of RNA dependent RNA polymerase (RdRp).

DISCUSSION AND CONCLUSION

A possible "micro" RNA complementary molecule to prevent the -1 frameshift is AGCAAAAAUUUGCCCAAACGCCAC. It binds in a complementary way to the genomic section that creates the FLNRVC as part of the RdRp enzyme. Suppose we construe a polymer that contains: (AGCAAAAAUUUGCCCAAACGCCAC)n with n>1. Then there is a statistical probability that one of the n AGCAAAAAUUUGCCCAAACGCCAC will meet the complementary group in the mRNA of ORF1ab of SARS-CoV-2 that makes the -1 frameshift. If such a proper repressor of UCGUUUUUAAACGGGUUUGCGGUG, is found like with e.g.

(*) 5'-ATG- (AGCAAAAAUUUGCCCAAACGCCAC)n-poly(A)-3'

Then the shift-1 will be prevented in the mRNA of de novo synthesis SARS-CoV-2 virus. Perhaps that the start signal 5'-ATG must be absent to avoid protein synthesis. Perhaps that protein repressors similar to MAF1, will be more effective to specifically repress the viral ORF1ab frameshift.

The consequence will be that the ORF1ab of that de novo virus is broken down and important enzymes for infection are not synthesized. This appears to be an interesting more medicinal approach to prevent COVID-19.

Note that the protein derived from ATGAGCAAAAAUUUGCCCAAACGCCAC, is MSKNLPKRHKK@. If, on the other hand, the RNA under (*) is somehow used as negative template [3] for positive RNA, then a number of proteins such as (SFLNGFAV)n can be synthesized as well. This latter protein can function as a biomarker to the suppressor RNA (*). It can be a quantitative measure for the blocking of frameshift-1 ORF1ab with the use of RNA (*). The more RNA (*) that vdWaals binds complementary to the ORF1ab genome that makes FLNRVC, the less (SFLNGFAV)n will be synthesized. Therefore the proposed concept of RNA repressor appears to be open to experimental cellular research.

Our *in silico* study showed that there are vulnerabilities and frameshift "mysteries" in the synthesis of essential proteins in the propagation of SARS-CoV-2 associated to the ORF1ab. We think that we also have delineated the contours of experiments that could lead to an mRNA type of medicinal treatment of COVID-19 or similar viral diseases. Especially it is noted that the first encountered stem in the frameshift effective pseudoknot is most likely not stable.

REFERENCES

1. Enjuanes, L. Preface to the bundle, Coronavirus replication and reversed engineering, (Edn) Enjuanes L, Springer Verlag Berlin Heidelberg. 2005.

2. Huang, X, Cheng, Q, Du, Z. A genome-wide analysis of RNA pseudoknots that stimulate efficient -1 ribosomal frameshifting or readthrough in animal viruses. BioMed Res Int. 2013;13(1):1-16.

3. Sawicki SG, Sawicki DL. Coronavirus transcription: A perspective. In coronavirus replication and reversed engineering, (Eds) Enjuanes L, Springer Verlag Berlin Heidelberg. 2005;287(16): 31-56.

4. Giedroc GP, Theimer CA, Nixon PL. Structure, stability and function of RNA Pseudoknots involved in stimulating ribosomal Frameshifting. J Mol Biol. 2000;298(47):167-185.

5. Staple DW, Butcher SE. Pseudoknots: RNA structures with diverse functions. PLoS Biol. 2005;8(7):0956-0958.

6. Gingras BL, Charbonneau J, Butcher SE. Targeting frameshifting in the human immunodeficiency virus. Expert Opin Ther Targets. 2012;16(6):249-258.

7. Gilbert W. Functional specialization of ribosomes?. Trends Biochem Sci. 2011;36(7):127-132.