Unheeded SARS-CoV-2 protein? Look deep into negative-sense RNA

Martin Bartas[1, *], Adriana Volná[2], Václav Brázda[3], Jiří Červeň[1] & Petr Pečinka[1, *]

[1]Department of Biology and Ecology, University of Ostrava, Ostrava, 710 00, Czech Republic

[2]Department of Physics, Faculty of Science, University of Ostrava, Ostrava, 71000, Czech Republic

[3]Institute of Biophysics, Czech Academy of Sciences, Brno, 612 65, Czech Republic

[*]Correspondence and requests for materials should be addressed to Martin Bartas (email: martin.bartas@osu.cz) or Petr Pečinka (email: petr.pecinka@osu.cz)

Abstract:

SARS-CoV-2 is a novel ssRNA+ virus from the *Coronaviridae* family, which has caused the global COVID-19 pandemic. The genome of SARS-CoV-2 is one of the largest of RNA viruses, comprising of 26 known protein-coding loci. This study aimed to explore the coding potential of negative-strand RNA intermediate for its potential to contain additional protein coding-loci. Surprisingly, we have found several putative ORFs and one brandt new functional SARS-CoV-2 protein-coding loci and called it *Avo1* (Ambient viral ORF1). This sequence is located on negative-sense RNA intermediate and *bona fide* coding for 81 amino acid residues long protein and contains strong Kozak sequence for translation on eukaryotic ribosomes. *In silico* translated protein Avo1 has a predominantly alpha-helical structure. The existence of *Avo1* gene is supported also by its evolutionarily and structural conservation in RaTG13 bat coronavirus. The nucleotide sequence of *Avo1* also contains a unique SREBP2 binding site which is closely related to the so-called "cytokine storm" in severe COVID-19 patients. Altogether, our results suggest the existence of still undescribed SARS-CoV-2 protein, which may play an important role in the viral lifecycle and COVID-19 pathogenesis.

Introduction:

Coronavirus SARS-CoV-2, as a causative agent of COVID-19 disease, is intensively studied worldwide. It is generally considered that the SARS-CoV-2 genome is coding for 16 non-structural proteins (NSP1-16), 4 structural proteins (surface glycoprotein, membrane glycoprotein, envelope protein, nucleocapsid phosphoprotein), and 6 – 9 "accessory factors" (still designated as "open reading frames", ORFs) (Gordon et al. 2020). The length of the SARS-CoV-2 proteins varied from 1945 aa to 13 aa (Huang et al. 2020; Li et al. 2020; Helmy et al. 2020). The shortest SARS-CoV-2 protein has only 13 aa (nsp11) and its function is unknown (Chen, Liu, a Guo 2020), followed by ORF10 (38 aa), ORF7b (43 aa), ORF6 (61 aa), and envelope protein (75 aa) (Wu et al. 2020). The longest SARS-CoV-2 protein is a multi-domain protein nsp3 (1945 aa) (Helmy et al. 2020). All to date described SARS-CoV-2 proteins are coded by positive-sense viral RNA strand. Overall, positive-sense RNA viruses are generally accepted to encode proteins solely on the positive strand, but in the current work by Dinan et al., they demonstrated that also negative-sense viral RNA strand has the potential to be protein-coding (Dinan et al. 2020). We have decided to inspect the negative-sense intermediate viral RNA strand of SARS-CoV-2 and look for new potential ORFs. As an indicator for *in vivo* validity, we have chosen a combination of complementary bioinformatics approaches, e.g. the presence of so-called Kozak sequence, which is considered strong supportive evidence for being translated on host ribosomes (Gong et al. 2014), transcription factor binding site analysis, homology search, and 3D structure prediction. Our analyses propose a novel SARS-CoV-2 ORF, which has a high probability to be translated in host cells.

Methods:

The SARS-CoV-2 reference genome (NC_045512.2) was searched by NCBI ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/), to predict all potential ORFs longer than 200 nucleotides. We have focused especially on the negative strand intermediate (i.e. frames -1, -2, and -3). Using ATG^{pr} web-based tool (https://atgpr.dbcls.jp/) (Salamov, Nishikawa, a Swindells 1998), we have inspected, if they contain so-called Kozak sequence, which is considered an important signal in the 5´ end of mRNAs to be translated on eukaryotic ribosomes (Noderer et al. 2014). To compute Mw and isoelectric point, we used the Expasy Compute pI/Mw tool (https://web.expasy.org/compute_pi/) (Gasteiger et al. 2005). For the transcription factor binding site analysis, the RegRNA 2.0. webserver was used (Chang et al. 2013). To find possible domain homologs we used NCBI´s Conserved Domains Database (CDD) webserver (Marchler-Bauer et al. 2015) with an E-value cut-off set to 10. For *ab initio* structural modeling, the QUARK tool was used (Xu a Zhang 2012) and the resulting structures were visualized within UCSC Chimera 1.15 workflow (Pettersen et al. 2004). Supplementary materials contain all nucleotide and predicted protein sequences (SM1). Predicted PDB structures are enclosed in SM2. CDD domain hits are included in SM3.

Results and discussion:

We have predicted all putative SARS-CoV-2 ORFs longer than 200 nucleotides, resulting in a set of 8 protein candidates (all ORFs are enclosed in SM1, the longest was 303 nt long, and the shortest 201 nt in size). Then we have inspected if they contain the so-called Kozak sequence. We have found one highly promising candidate locus and called it *Avo1* (Ambient viral ORF1). *Avo1* nucleotide sequence (Figure 1 Top) contains a very well-conserved Kozak signal in the form of RYMRMVAUGGC (Noderer et al. 2014), which is a nucleic acid motif that functions as the protein translation initiation site in most mammalian mRNA transcripts (Noderer et al. 2014). The strongest Kozak signals found in the SARS-CoV-2 genome are depicted in Figure 1 Bottom (five of six Kozak signals are located in RNA+ strand, one is located on the RNA- strand).

Avo1 locus is situated on nucleotide positions 422 – 667 of the reference SARS-CoV-2 genome (NC_045512.2) and putative protein is translated in reading frame -2 (i.e. only from negative-sense SARS-CoV-2 RNA intermediate). Protein product would be 81 amino acid residues long and its molecular weight was computed to be 9586.06 Da with an isoelectric point (pI) 9.22. Interestingly, the proposed *Avo1* gene contains 12 nt long SREBP2 binding site (Figure 1 Top). Recently it was found that "COVID-19-activated SREBP2 disturbs cholesterol biosynthesis and leads to cytokine storm" (Lee et al. 2020).

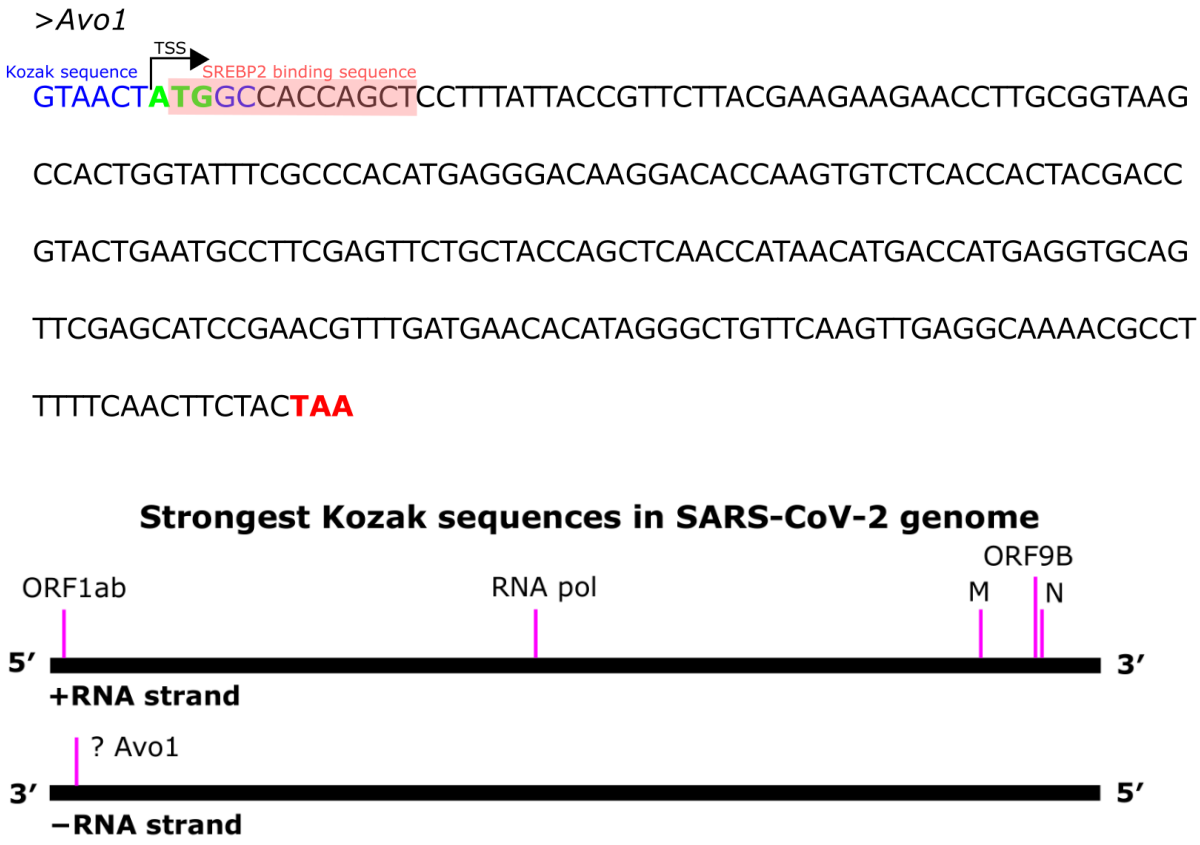

Figure 1. (Top) Genomic sequence of putative *Avo1* gene. Kozak sequence is depicted in blue color, start codon ATG in green, stop codon TAA in red. SREBP2 binding sequence is highlighted by pink. TSS stands rather for "translation start site", because we are dealing already with viral RNA. (Bottom) Six highly significant Kozak sequences in the SARS-CoV-2 genome, the *Avo1* signal is the second strongest one (see SM4).

Homology searches revealed that *Avo1* ORFs are evolutionarily conserved in several coronaviruses, e.g. in bat coronaviruses RaTG13, RpShaanxi2011, BtRs-BetaCoV/HuB2013, Rc-o319, or in pangolin coronavirus GX/P2V (Figure 2 Top). On the other side, human SARS-CoV and MERS-CoV *Avo1* loci are truncated by stop codons - dynamic loss or gain of specific accessory proteins may be an explanation for different viral properties (virulence, pathogenesis, etc.). Three weak domain homology signals were found in CDD database: B-12 dependent methionine synthase, Herpes envelope superfamily, and adenosine deaminase.

Another evidence for possible Avo1 protein biological significance is that instead of some disordered clumps, an independent *ab initio* prediction of tertiary structures of SARS-CoV-2 Avo1 and RatG13 Avo1 has shown five alpha-helices forming conserved structures (Figure 2 Bottom).

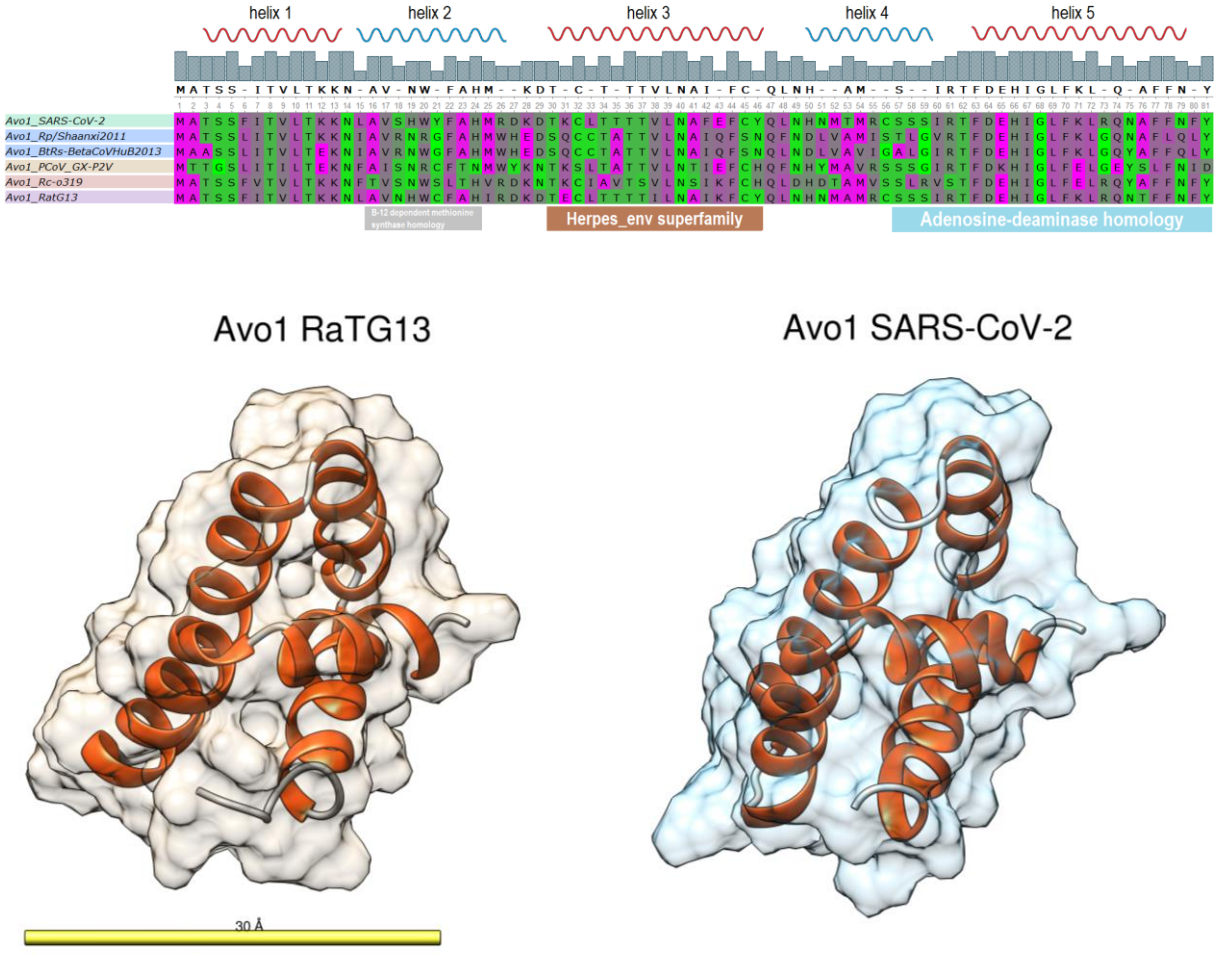

Figure 2. (Top) Alignment of Avo1 putative proteins from homologous coronaviruses. Coloring express helix propensity, magenta – high, green – low. Consensus sequence inferred from the alignment bar plotted. Domain hits for the SARS-CoV-2 sequence are inside the rectangles. (Bottom) *De novo* modeled Avo1 structure using QUARK workflow (Xu and Zhang 2012), yellow scale bar is 30 Å long.

It was found that the SARS-CoV-2 transcriptome is very diverse and rich (Kim et al. 2020) and therefore, it is likely that also SARS-CoV-2 proteome will follow this trend and still unheeded viral proteins or protein variants wait for its discovery. Every newly described SARS-CoV-2 protein (and its host interacting proteins) could be a potential target for a drug repurposing or development of novel therapies (Gordon et al. 2020), for instance it was found, that SARS-CoV-2 accessory proteins ORF6 and ORF8 inhibit type I interferon signaling pathway, and thus play a critical role in innate immune suppression during viral infection (Li et al. 2020). Altogether, our results suggest the existence of a still undescribed SARS-CoV-2 protein, which may play an important role in the viral lifecycle and COVID-19 pathogenesis.

Acknowledgment: We would like to thank our institution, the University of Ostrava, for an inspiring working environment and academic freedom.

Author contribution:

M.B. carried out most of the analyses and wrote the manuscript, A.V. has conceived the presented idea and help with data curation, V.B. and J.C. have contributed to writing and revisions, P.P. supervised the work.

References:

Dinan, Adam M., Nina I. Lukhovitskaya, Ingrida Olendraite, a Andrew E. Firth. 2020. „A Case for a Negative-Strand Coding Sequence in a Group of Positive-Sense RNA Viruses". *Virus Evolution* 6 (1). https://doi.org/10.1093/ve/veaa007.

Gasteiger, Elisabeth, Christine Hoogland, Alexandre Gattiker, Marc R. Wilkins, Ron D. Appel, a Amos Bairoch. 2005. „Protein identification and analysis tools on the ExPASy server". In *The Proteomics Protocols Handbook*, 571–607. Springer.

Gong, Yu-Nong, Guang-Wu Chen, Chi-Jene Chen, Rei-Lin Kuo, a Shin-Ru Shih. 2014. „Computational Analysis and Mapping of Novel Open Reading Frames in Influenza A Viruses". *PLOS One* 9 (12): e115016. https://doi.org/10.1371/journal.pone.0115016.

Gordon, David E., Gwendolyn M. Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M. White, Matthew J. O'Meara, et al. 2020. „A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug Repurposing". *Nature* 583 (7816): 459–68. https://doi.org/10.1038/s41586-020-2286-9.

Helmy, Yosra A., Mohamed Fawzy, Ahmed Elaswad, Ahmed Sobieh, Scott P. Kenney, a Awad A. Shehata. 2020. „The COVID-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control". *Journal of Clinical Medicine* 9 (4): 1225.

Huang, Yuan, Chan Yang, Xin-feng Xu, Wei Xu, a Shu-wen Liu. 2020. „Structural and Functional Properties of SARS-CoV-2 Spike Protein: Potential Antivirus Drug Development for COVID-19". *Acta Pharmacologica Sinica* 41 (9): 1141–49. https://doi.org/10.1038/s41401-020-0485-4.

Chang, Tzu-Hao, Hsi-Yuan Huang, Justin Bo-Kai Hsu, Shun-Long Weng, Jorng-Tzong Horng, a Hsien-Da Huang. 2013. „An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs". In *BMC Bioinformatics*, 14:1–8. BioMed Central.

Chen, Yu, Qianyun Liu, a Deyin Guo. 2020. „Emerging Coronaviruses: Genome Structure, Replication, and Pathogenesis". *Journal of Medical Virology* 92 (4): 418–23. https://doi.org/10.1002/jmv.25681.

Kim, Dongwan, Joo-Yeon Lee, Jeong-Sun Yang, Jun Won Kim, V. Narry Kim, a Hyeshik Chang. 2020. „The Architecture of SARS-CoV-2 Transcriptome". *Cell* 181 (4): 914-921.e10. https://doi.org/10.1016/j.cell.2020.04.011.

Lee, Wonhwa, June Hong Ahn, Hee Ho Park, Hong Nam Kim, Hyelim Kim, Youngbum Yoo, Hyosoo Shin, Kyung Soo Hong, Jong Geol Jang, a Chun Gwon Park. 2020. „COVID-19-activated SREBP2 disturbs cholesterol biosynthesis and leads to cytokine storm". *Signal Transduction and Targeted Therapy* 5 (1): 1–11.

Li, Jin-Yan, Ce-Heng Liao, Qiong Wang, Yong-Jun Tan, Rui Luo, Ye Qiu, a Xing-Yi Ge. 2020. „The ORF6, ORF8 and Nucleocapsid Proteins of SARS-CoV-2 Inhibit Type I Interferon Signaling Pathway". *Virus Research* 286: 198074. https://doi.org/10.1016/j.virusres.2020.198074.

Marchler-Bauer, Aron, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y. Geer, Renata C. Geer, Jane He, Marc Gwadz, a David I. Hurwitz. 2015. „CDD: NCBI's conserved domain database". *Nucleic Acids Research* 43 (D1): D222–D226.

Noderer, William L., Ross J. Flockhart, Aparna Bhaduri, Alexander J. Diaz de Arce, Jiajing Zhang, Paul A. Khavari, a Clifford L. Wang. 2014. „Quantitative analysis of mammalian translation initiation sites by FACS-seq". *Molecular Systems Biology* 10 (8): 748.

Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, a Thomas E. Ferrin. 2004. „UCSF Chimera—a visualization system for exploratory research and analysis". *Journal of Computational Chemistry* 25 (13): 1605–1612.

Salamov, A. A., T. Nishikawa, a M. B. Swindells. 1998. „Assessing Protein Coding Region Integrity in CDNA Sequencing Projects". *Bioinformatics (Oxford, England)* 14 (5): 384–90. https://doi.org/10.1093/bioinformatics/14.5.384.

Wu, Fan, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, a Yuan-Yuan Pei. 2020. „A new coronavirus associated with human respiratory disease in China". *Nature* 579 (7798): 265–269.

Xu, Dong, a Yang Zhang. 2012. „Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field". *Proteins: Structure, Function, and Bioinformatics* 80 (7): 1715–1735.