BIOCHEMISTRY, BIOPHYSICS, AND MOLECULAR BIOLOGY

Novel Negative Sense Genes in the RNA Genome of Coronaviruses

Corresponding Member of the RAS O. P. Zhirnov[a,b,*] and S. V. Poyarkov[a]

Abstract—The coronavirus family consists of lipid-containing envelope viruses that have a single-stranded RNA genome that encodes 25—30 proteins in different viruses by the mechanism of positive-polarity strategy. In addition, extended open reading trnslation frames (ORFs, genes) located in a negative-sense orientation were found in the genomes of coronaviruses. The size of negative-sense genes varies in the range of 150—450 nt, which corresponds to polypeptides encoded by negative-polarity genes (negative gene proteins, NGP) with m. m. $5-30 \times 10^3$ kDa. Coronaviruses show marked differences from virus to virus in the number of negative genes detected. These negative-sense genes in the coronavirus genome allow this family to be considered as viruses developing an ambisense genome strategy.

Coronaviruses are of great interest due to the threatening development of a human pandemic caused by the SARS-CoV2 strain [1, 2]. Viruses of this family have an outer lipid envelope (enveloped viruses) and a single-stranded genomic RNA, which has a classic positive polarity and, in different viruses, encodes 25—30 specific proteins [3]. Expression of viral proteins occurs in two phases. In the first phase, the 5'-terminal part of the genome is translated with the formation of 19 nonstructural proteins, ns1—ns19 (Fig. 1a). In the second phase, the 3'-terminal part of the genome is translated through intermediate replication and transcription of the genome with the formation of individual mRNAs encoding structural proteins N, M, S, E, HE, and a number of nonstructural accessory proteins (Fig. 1a).

As a result of in silico analysis of the coronaviral genomic RNA, we found extended open reading frames (ORFs) that started with the AUG codon and ended with the classical termination codons UAA or UAG (Table 1, Fig. 1b). Analysis of the negative-polarity zones preceding the AUG position in the identified genes, in particular, the largest gene NGP4 (the negative gene protein 4) (nc 6137–6489) in the SARS-CoV2 virus genome, which was performed using a computer prediction program for ribosome-binding elements, showed the presence of regions with

a pronounced secondary structure with numerous hair-pins and, according to the criteria HTTP://bio-info.net.in/IRESPred software [5], exhibiting a high energy stability (free energy 85 kcal/mmol) and structural properties of the internal ribosome entry site (IRES) (Fig. 1c). Such a structure of the 5'-ATG-adjacent zone can provide recognition of mRNA by ribosomes and subsequent translation of the protein [5]. Moreover, in this IRES-like zone, two additional AUGs and three alternative CUG initiation codons in the translation phase +1 and/or +2 were detected, which might also facilitate the recognition and expression of this gene by ribosomes by the scanning mechanism [4].

The length of the detected genes (ORF) varied in the range of 150—450 nucleotides (nt), which could ensure the synthesis of polypeptides from the molecular weight of at 5—30 kDa. Comparison of the genomes of various members of the coronavirus family showed a significant diversity both in the number of such negative-polarity genes and in the pattern of their localization in the viral genome (Table 1, Fig. 1b). For example, the pangolin-CoV2 and SARS-CoV2 viruses, which, according to modern concepts, are the closest relatives (i.e., generations of the same predecessor), were shown to contain 29 and 21 negative genes, respectively, in the absence of coincidence of their positions in the genome (Fig. 1b). In contrast, the comparison of the BAT-CoV and SARS-CoV2 viruses, belonging to the same genus of beta coronaviruses, showed that they have a similar number of classic AUG-containing negative genes (17 and 21, respectively), which, moreover, have a similar localization in the genome. Thus, the presence and similar localization of these genes in the genome of human

a Russian—German Academy of Medico-Social and Biotechnological Sciences, Moscow, Russia

b Ivanovsky Institute of Virology, Gamaleya National Research Center of Epidemiology and Microbiology, Moscow, Russia
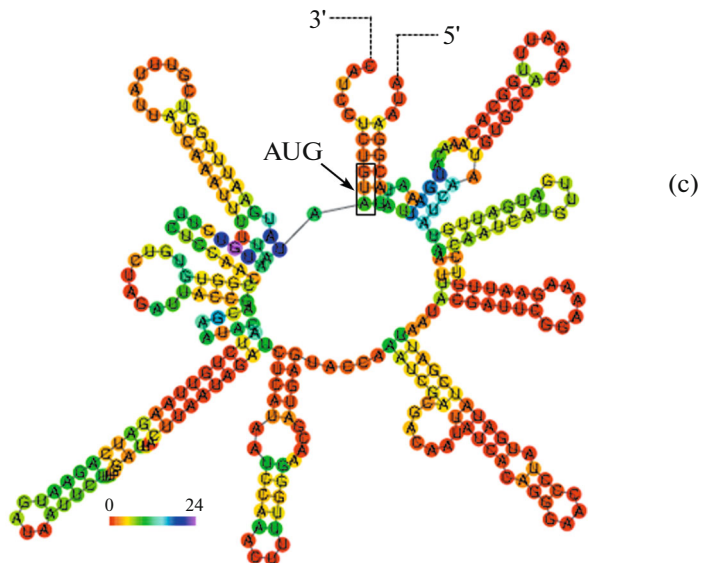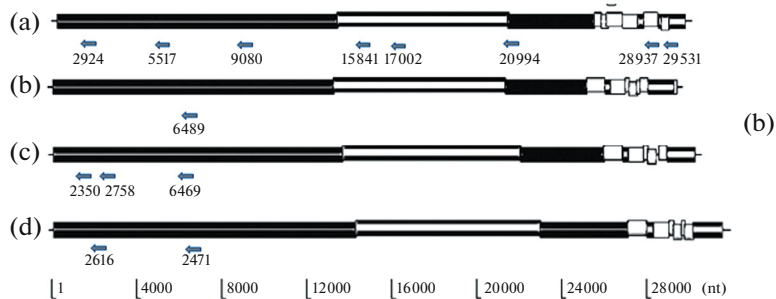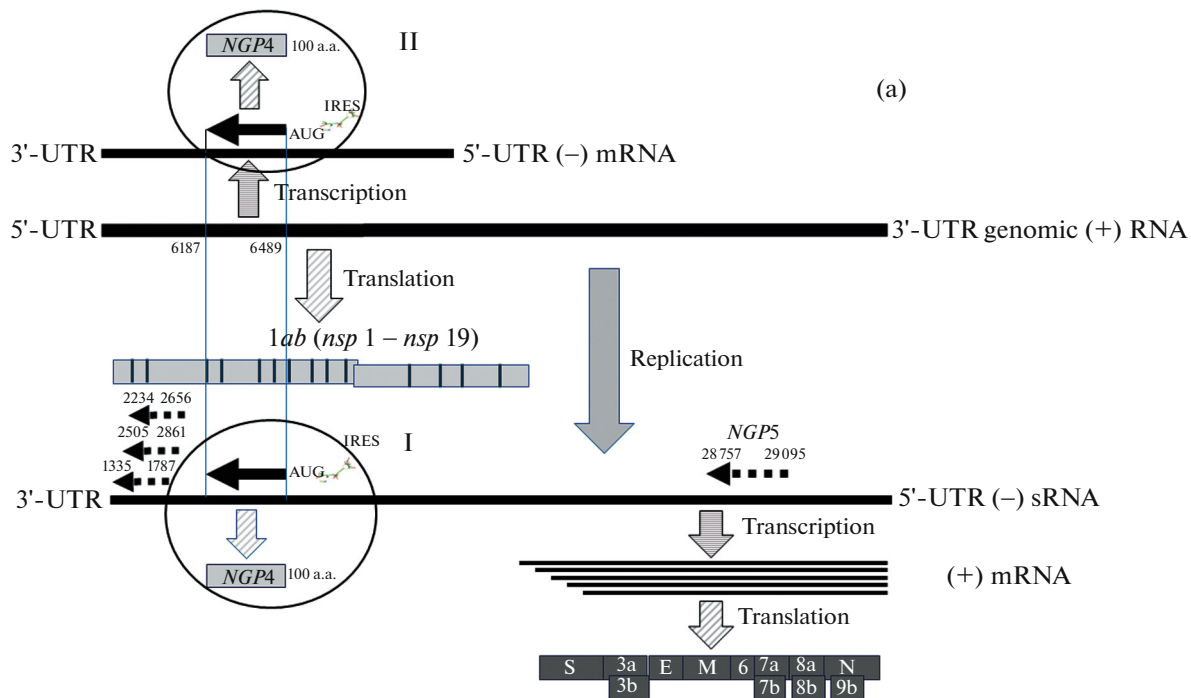
*e-mail: zhirnov@inbox.ru

(a)

(b)

(c)

Table 1. Quantitative characteristics of the detection of negative-polarity genes (ORF) in the genomes of coronaviruses

Viruses [1]	Number of small negative-polarity genes with ATG/CUG codons 150—300 nt long [2]	Number of large negative-polarity genes with AUG/CUG codons ≥300 nt long	Molecular mass range for large polypeptides (NGP) (kDa × 10^{-3}) [3]
Alpha coronaviruses human coronavirus HCoV-229E	29/29	1/5	12.37—14.41
Beta-coronaviruses			
SARS-CoV	34/35	0/2	11.47—14.98
MERS-CoV	32/23	8/3	11.07—18.55
SARS-CoV2	21/26	1/4	10.85—17.18
Pangolin-CoV	29/17	3/4	10.8—19.9
BAT-CoV	2/1	17/29	10.95—19.27
HKU1	15/13	1/2	13.1—21.8
Gamma-coronaviruses Avian infectious bronchitis virus	20/8	6/3	12.73—26.46
Delta-coronaviruses Porcine coronavirus HKU15	26/29	5/3	11.19—17.35

(1) The number of negative-polarity genes (ORFs, open reading frames) and the polypeptides encoded by them (negative genes proteins, NGPs) are shown for each virus. The analysis was performed on the basis of the primary structure of the genes of the specified viruses: HCoV-229E (ac.n. NC_002645.1), MERS-CoV (NC_019843.3), SARS-CoV (NC_004718.3), SARS-CoV2 (MT635445.1), Pangolin-CoV (MT040335.1), avian infectious bronchitis virus (NC_001451.1), porcine coronavirus HKU-15 (NC_039208.1), bat coronavirus Bat/RATG13 (MN996532.1), and human coronavirus HKU1 (NC_006577.2).

(2) Number of open reading frames in the negative polarity of the genome, which start with the classic initiator codon AUG (left numeral) and, next to slanting line, with the alternative CUG codon (right numeral) and end with a termination codon (UAA, UAG, or UGA) [4].

(3) Ranges of molecular weights of polypeptides from the group of negative genes with a length of ≥ 300 nt, which were calculated on the basis of the primary gene sequence.

viruses and bats confirms the genetic and evolutionary proximity of these viruses.

Conversely, the identification of 29 AUG-negative genes in the genome of the Pangolin-Cov virus (Table 1) may indicate that, contrary to modern concepts, the virus from pangolins is a more distant relative of SARS-CoV2 than the bat virus.

Additional extended ORFs were detected in the genome of coronaviruses if the alternative initiation codon CUG was used as the start codon (Table 1). Similarly to the ORF with the classic AUG, the alternative-type ORFs had IRES-like structures and could provide the synthesis of extended polypeptides with molecular weights in the range of 5—30 kDa. The presence of additional negative-polarity genes of an alternative type in the genome of coronaviruses can significantly increase its genetic capacity.

The results of this report show the presence of extended reading frames (genes) in the genome of coronaviruses, the peculiarity of which is that these genes have a negative orientation. At the same time, the genome of coronaviruses is currently considered to be positive-polar, since all known genes of coronaviruses (approximately 25 genes for the nonstructural proteins and 5 major genes for the structural proteins (E, M, S, N, and HE)) are encoded in genomic RNA

Fig. 1. Replication scheme of the RNA genome and localization of negative-polarity genes in the genome of coronaviruses. (a) Scheme of the genomic viral RNA and the reading frame of viral genes based on the coronavirus model (ac.n. MT890462.1; SARS-Cov2/human/RUS/20200417_10/2020). The vertical shift of the reading frame in the viral gene reflects the shift in the translation phase. UTR—untranslated terminal regions in RNA. The positive-polarity nonstructural (nsp1-nsp19, 3a, 4c, 7ab, 8a, 9b) and structural (N, S, M, E, HE) genes are shown with gray boxes. The negative-polarity genes and proteins are designated NGP (negative genes proteins) and are shown with wide arrows shaded with a fine grid. Proteins NGP1 (nt positions 1335—1787), NGP2 (2505—2861), NGP3 (2234—2656), and NGP5 (28757—29095) begin with the dominant alternative CUG codon (dashed arrows); the NGP4 protein (nt 6187—6489; the zone of the "positive-polarity" protein nsp3 (Mpro)) begins with the classic AUG codon (solid arrow). (b) Schemes of localization of large negative-polarity genes (≥300 nt) of four coronaviruses: (a) MERS-CoV, (b) SARS-CoV2, (c) Pangolin-CoV, and (d) Bat SARS-like CoV RatGT13. The schemes are based on the analysis of the nucleotide sequence in the GenBank database: human coronavirus MERS (NC_019843.3), SARS-CoV (NC_004718.3), SARS-CoV2 (MT635445.1), Pangolin-CoV (MT040335.1), and Bat/RATG13 (MN996532.1). The numbers indicate the nucleotide at the beginning of the gene, counting from the 5'-end of the viral genome. (c) Scheme of the secondary structure of the region upstream of the NGP4 gene (positions 6187—6489 nt from the 5'-end of genomic (+)RNA) obtained using the *HTTP://bioinfo.net.in/IRE-SPred* software [5]; the arrow indicates the AUG initiation codon.

in a positive orientation and have an appropriate strategy of genome expression in infected target cells (Fig. 1a). The presence of new negative-polarity genes implies the existence of two mechanisms of their expression and possible synthesis of the corresponding mRNAs and subsequent translation of proteins in two possible ways: either direct translation of a replicative copy of genomic (−)RNA (pathway I) or the transcription of genomic (+)RNA with the formation of individual mRNAs of "negative polarity" for their subsequent translation with the formation of specific polypeptides (pathway II) (Fig. 1a, circled). Interestingly, in the genome of another family, influenza viruses belonging to the family of orthomyxoviruses, which are characterized by a negative-polarity strategy of genomic RNA, (ambisense) positive-polarity genes encoded on the viral negative-polarity genome were detected in a similar way [6−10].

The function and role of the newly discovered ambipolar viral genes have not yet been established. In the case of influenza viruses, there is an assumption that the identified new ambisense genes may be important in the regulation of the immune response to viral proteins and/or in the regulation of the stability of viral proteins in infected cells through the protein deubiquitination system [11−14]. To understand the possible functional significance of the identified new ambipolar genes, it is necessity to take into account two features inherent in these genes. First, the evolutionary stability of the existence of ambipolar genes in viruses for a long time indicates their biological determination [11]. Second, the coding of genes with opposite polarity in the same region of the RNA molecule in the so-called genes stacking format makes it possible to significantly increase the genetic capacity of the viral genome and opens up new opportunities for the virus for variability, increasing adaptability to the host, and biological evolution in nature [11].

The presence of multiple ambisence genes opens up a real possibility of coding a multivirionic population consisting of virions of different structural types, when more than one type of viral particles with an identical genome but a different composition of structural proteins can be synthesized from one genome. In this case, part of the virions (possibly infectious) may remain invisible (the principle of the "dark side of the Moon"). Moreover, this multivirionic profile of the virus population, programmed by the viral genome, may have a cellular or tissue dependence, in which each type of viral particles will have autonomous replication and reproduction and dominate in a particular host (organ or tissue). This as yet hypothetical phenomenon of replication of multiviral particles on the same genome may be important in the cell- or organ-dependent pathogenesis of a viral disease and may create new platforms for the development of methods of treatment and vaccine prevention.

The discovered new negative-polarity genes in the genome of coronaviruses have specific localization for the viral strain and quantitative composition in the genome (Fig. 1b). Thus, the pattern of negative-polarity genes in the viral strain genome can serve as its molecular signature and be used in the diagnosis and study of viral relationships and biological evolution of the coronavirus family.

The presence of potential negative-polarity genes in the genome of coronaviruses raises the question of the classification of this family. The detection in infected cells or infected organisms of protein products expressed on the "negative" gene template gives grounds for classifying the coronavirus family with the ambisens viruses with a bipolar genome strategy.

Currently, such ambisens viruses include viruses of four genera: phlebo-, tospo-, arena-, and tenuviruses [15]. The ambisense genes located in the genome in the stacking format were found in influenza viruses, in which, similarly to coronaviruses, direct expression of these genes has not yet been identified, but there are indirect signs of such expression during natural viral infection in vivo [12, 13]. The study of the mechanisms of the possible expression of the genetic information of these new genes, as well as the elucidation of the role and significance of the detected genes and/or their protein products during viral replication can serve as the basis for creating a new type of vaccines and antiviral chemotherapy agents for treatment of coronavirus infection.

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that they have no conflict of interest. This article does not contain any studies involving animals or human participants performed by any of the authors.

REFERENCES

1. Wu, F., Zhao, S., Yu, B., et al., A new coronavirus associated with human respiratory disease in China, *Nature*, 2020, vol. 579, pp. 265−269.
   https://doi.org/10.1038/s41586-020-2008-3
2. Guo, Y.R., Cao, Q.D., Hong, Z.S., Tan, Y.Y., Chen, S.D., Jin, H.J., Tan, K.S., Wang, D.Y., and Yan, Y., The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak—an update on the status, *Mil. Med. Res.*, 2020, vol. 7, no. 1, p. 11.
   https://doi.org/10.1186/s40779-020-00240-032169119
3. Fehr, R.A. and Perlman, S., Coronaviruses: an overview of their replication and pathogenesis, *Methods*

Mol. Biol., 2015, vol. 1282, pp. 1−23.
https://doi.org/10.1007/978-1-4939-2438-7_1

4. Kearse, M.G. and Wilusz, J.E., Non-AUG translation: a new start for protein synthesis in eukaryotes, *Genes Dev.*, 2017, vol. 31, no. 17, pp. 1717−1731.
https://doi.org/10.1101/gad.305250.117

5. Kolekar, P., Pataskar, A., Kulkarni-Kale, U., Pal, J., and Kulkarni, A., IRESPred: Web server for prediction of cellular and viral internal ribosome entry site (IRES), *Sci. Rep.*, 2016, vol. 6, p. 27436.
https://doi.org/10.1038/srep2743627264539

6. Palese Baez, M., Zazra, J.J., Elliott, R.M., Young, J.F., and Palese, P., Nucleotide sequence of the influenza A/duck/Alberta/60/76 virus NS RNA: conservation of the NS1/NS2 overlapping gene structure in a divergent influenza virus RNA segment, *Virology*, 1981, vol. 113, no. 1, pp. 397−402. PMID: 6927848

7. Zhirnov, O.P., Poyarkov, S.V., Vorob'eva, I.V., Safonova, O.A., Malyshev, N.A., and Klenk, H.D., Segment NS of influenza A virus contains an additional gene NSP in positive-sense orientation, *Dokl. Biochem. Biophys.*, 2007, vol. 414, pp. 127−133. PMID: 17695319

8. Zhirnov, O.P., Vorobjeva, I.V., Saphonova, O.A., Poyarkov, S.V., Ovcharenko, A.V., Anhlan, D., and Malyshev, N.A., Structural and evolutionary characteristics of *HA, NA, NS* and *M* genes of clinical influenza A/H3N2 viruses passaged in human and canine cells, *J. Clin. Virol.*, 2009, vol. 45, no. 4, pp. 322−333.
https://doi.org/10.1016/j.jcv.2009.05.03019546028

9. Zhirnov, O.P., Akulich, K.A., Lipatova, A.V., and Usachev, E.V., Negative-sense virion RNA of segment 8 (NS) of influenza A virus is able to translate in vitro a new viral protein, *Dokl. Biochem. Biophys.*, 2017, vol. 473, no. 1, pp. 122−127.
https://doi.org/10.1134/S160767291702009028510127

10. Gong, Y.N., Chen, G.W., Chen, C.J., Kuo, R.L., and Shih, S.R., Computational analysis and mapping of novel open reading frames in influenza A viruses, *PLoS One*, 2014, vol. 9, no. 12. e115016.
https://doi.org/10.1371/journal.pone.011501625506939

11. Zhirnov, O.P., Unique bipolar gene architecture in the RNA genome of influenza A virus, *Biochemistry* (Moscow), 2020, vol. 85, no. 3, pp. 387−392.
https://doi.org/10.1134/S000629792003014132564743

12. Hickman, H.D., Mays, J.W., Gibbs, J., Kosik, I., Magadan, J.G., Takeda, K., and Yewdell, J.W., Influenza A virus negative strand RNA is translated for CD8+ T cell immunosurveillance, *J. Immunol.*, 2018, vol. 201, pp. 1222−1228.
https://doi.org/10.4049/jimmunol.1800586

13. Zhirnov, O.P., Konakova, T.E., Anhlan, D., Ludwig, S., and Isaeva, E.I., Cellular immune response in infected mice to nsp protein encoded by the negative strand NS RNA of influenza A virus, *MIR J.*, 2003, vol. 6, no. 1, pp. 28−36.
https://doi.org/10.18527/2500-2236-2019-6-1-28-36

14. Zhong, W., Reche, P.A., Lai, C.C., Reinhold, B., and Reinherz, E.L., Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire, *J. Biol. Chem.*, 2003, vol. 278, no. 46, pp. 45135−45144. PMID: 12960169

15. Nguyen, M. and Haenni, A.L., Expression strategies of ambisense viruses, *Virus Res.*, 2003, vol. 93, no. 2, pp. 141−150. PMID: 12782362

Translated by M. Batrukova