Determinative Role of Non-structural Protein Genes in Genus Identification of Coronaviruses

Maria Chaley
Bioinformatics laboratory
Institute of mathematical problems of biology RAS – branch of Keldysh Institute of applied mathematics RAS
Pushchino, Russia
chaley@phystech.edu

Vladimir Kutyrkin
Department of Fundamental sciences
Moscow state technical university n.a. N.E. Bauman
Moscow, Russia
vkutyrkin@yandex.ru

2022 Ural-Siberian Conference on Computational Technologies in Cognitive Science, Genomics and Biomedicine (CSGB) | 978-1-6654-5288-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/CSGB56354.2022.9865492

***Abstract*— Interrelationship of coronavirus genus with key fragments of viral genome was investigated. Genes of structural proteins (S-gene of spike protein and N-gene of nucleocapsid protein) and ORF1ab of polyprotein pp1ab, that in infected cell is split into 16 non-structural proteins, were considered as such fragments. Statistical method based on averaged codon distribution in the genes of genus prototype variants was applied in the work to recognize genus of coronavirus. High reliability of this method has been demonstrated earlier in recognizing the 15 species and serotypes of the flaviviruses, such as viruses of yellow fever, dengue fever, various encephalitides, etc. For each key fragment of the coronavirus genome the numerical experiments on identification of genus for the 3242 viral genomes from the GenBank have been done. The highest reliability (98%) was achieved, when ORF1ab frequency codon characteristics were used. It appeared to be that in recognizing genus of *Gammacoronavirus*, basing on spike protein gene, about half of the 345 genomes of this genus were identified as *Betacoronavirus* (84.6%) and *Alphacoronavirus* (15.4%). Analogous phenomenon of significant error appeared in determinating *Alphacoronavirus* genus, basing on nucleocapsid protein gene, also. However, these significant errors may be a consequence of the coronavirus genome plasticity in the result of homologous recombinations between the viral genomes.**

Keywords—identification of coronavirus genus, ORF1ab, S-gene, N-gene

I. Introduction

Coronaviruses (CoVs) cause the diseases of various degree of severity in the birds, pigs, cattle, in the cats and dogs, camels, in the shrews, rats and mouse-like rodents, in the hedgehogs and bats, the whales, etc. At present time seven coronaviruses causing the diseases in humans are known. Usually, they are divided into the two groups: HCoV 229E, HCoV NL63, HCoV HKU1, HCoV OC43 – that predispose a disease progression generally like an ARVI (acute respiratory viral infection), and highly dangerous ones, such as SARS-CoV-1 (severe acute respiratory syndrome-related coronavirus), MERS-CoV (middle east respiratory syndrome coronavirus) and SARS-CoV-2 (severe acute respiratory syndrome 2 coronavirus) induced the COVID-19 pandemic.

Taxonomic classification of the coronaviruses has been drawn up by the year 2011, when in catalog of the International Committee on Taxonomy of Viruses (ICTV) status of *Coronavirus* genus was changed onto subfamily *Coronavirinae* including the four genera: *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus* and *Gammacoronavirus* [1]. Taxonomic structure of the coronaviruses has been essentially revised in the year 2018 in the result of introduction of a notion of subgenus and mathematical assessment of a few ranks (subgenus, genus, family) according to phylogenetic tree constructed by the likelihood method after multiply alignment of full-size genomes for all known family representatives [2]. Moreover, a new coronavirus discovered in the year 2018 by metagenomic analysis and named as letovirus 1 of narrow-mouthed toads (*Microhyla fissipes*) was singled out by ICTV into distinct subfamily *Letovirinae* [3], while other earlier known coronaviruses constituted *Orthocoronavirinae* subfamily [2].

Only coronaviruses of *Orthocoronavirinae* subfamily are under consideration in the present work. It is found that the subfamily is related to the volatiles which are a reservoir of the coronaviruses: chiroptera – for *Alphacoronavirus* (α–CoV) and *Betacoronavirus* (β–CoV) genera, birds – for *Deltacoronavirus* (δ–CoV) and *Gammacoronavirus* (γ–CoV) genera [4].

The coronaviruses assumed such name after their form similar to solar corona has been revealed with the help of electron microscopy [5]. Such the form is due to viral virion is edged with clublike spikes of surface glycoprotein S (spike-protein). Spike-protein is responsible for virus entry into cell by binding with certain transmembrane receptors. Two other structural proteins are anchored in lipid envelope of virion also: channel-forming E-protein (envelope protein) and M-protein (membrane protein). Virion genomic RNA$^+$ (vgRNA of positive polarity) and N-protein (nucleocapsid protein) are inside nucleocapsid [1, 6]. Genomic RNA$^+$ serves as mRNA to synthesize two long polyproteins pp1a and pp1ab of length about 4000 and 7000 amino acid residues, correspondingly. Polyprotein pp1ab, including pp1a, is formed in the result of ribosome ignoring the stop-signal due to a hairpin RNA loop that shifts reading frame at one nucleotide back, as a rule. Reading frames of such the polyproteins in the genome are designated as ORF1a and ORF1ab. In infected cells from the polyproteins pp1a and pp1ab two proteases are released: major protease Mpro and papain-like protease PLpro which cut the whole polyprotein pp1a into 11 non-structural polyproteins (nsp) and polyprotein pp1ab into 16 distinct non-structural proteins. These non-structural proteins carry out various important functions in coronavirus life cycle [7, 8]. For example, nsp12 is RNA-dependent RNA-polymerase (RdRp). Using vgRNA$^+$ as matrix, RdRp synthesizes complementary to genomic RNA of negative-sense (cgRNA$^-$) which acts as matrix to synthesize vgRNA$^+$ for new virions. Moreover, RdRp synthesizes a series of subgenomic RNAs of negative polarity (sgRNA$^-$) on the vgRNA matrix. Further such the sgRNAs of negative-sense are used for synthesis of subgenomic matrix RNAs of positive polarity (sgmRNA) from which some non-structural and structural proteins, particularly, S-, M- and N- proteins are translated. Coronaviruses have the longest genome among all RNA-viruses, that constitutes in the order of 30 000 nucleotides and the 2/3 of them are occupied by ORF1ab.

Natural variability of the coronavirus genome is provided with spontaneous mutations during its replication and homologous recombinations between other viral genomes [9, 10]. Coronaviruses accumulate less mutations than the majority of RNA-viruses because they encode ferment nsp14, correcting replication mistakes [11]. S-protein gene is hypervariable in the genome of coronaviruses and responsible for their interspecies transmission, virulence and contagiosity. Frequency of mutations in various coronavirus species is different. For example, in ACoVs, provoking respiratory infections of the birds, mutation frequency in the S-gene is by an order of magnitude greater ($3-6\times10^{-3}$ changes in year/site [12]) than in HCoVs 229E giving rise to the respiratory diseases in humans (3×10^{-4} changes in year/site [15]). However, the mutations span the more genes, than S-protein gene only.

Alignment of the genome sequences of coronaviruses has revealed 58 % of homology in area encoding non-structural proteins, 43 % – in region encoding structural proteins, and 54 % – at level of the whole genome, that allows considering non-structural proteins as more conservative, but structural proteins as more variable and supporting virus adaptation to novel hosts [13].

This is more than two years as pandemic given rise by SARS-CoV-2 coronavirus does not die down, going through new waves of mutant variants. Along with searching for the means to combat against the coronavirus a question of risk assessment for arising novel epidemic sources becomes very actual. The more various and accurate instruments will be applied to identify the coronaviruses and analyze them, the rather sooner and more successful such the assessment may be done. So, our work was aimed at development of fast and reliable method to determine coronavirus genus basing on the key fragments of the viral genome. Three fragments have been considered as the key ones: ORF1ab which encoding non-structural proteins, S-gene of spike-protein and N-gene of nucleocapsid protein. Procedure of coronavirus genus recognition, basing on statistical method proposed earlier, has been considered for each fragment individually. In applying the procedure, we rested on previous experience of development of statistical method for recognizing flavivirus species (among them there were viruses of yellow fever, dengue fever, West Nile fever, etc.), basing on the known genome sequence [14]. As shown in the work, the most reliability in coronavirus genus recognition was achieved, if ORF1ab has been used.

II. Materials and Methods

Earlier, in developing the method to recognize flavivirus species, a statistic basing on comparison of codon distribution in the genes of flavivirus polyproteins has been used [14]. In the present work the same statistics is applied for revealing determinative interrelationship between coronavirus genus and distinctly considered group of non-structural genes (ORF1ab) and the two structural genes (S- and N-gene).

For each of three genes (ORF1ab, S- and N-gene) in the genome sequence of prototype variant of coronavirus a distribution of codon frequency was considered. Quantities of analyzed genomes of prototype variants were the following: 22, 28, 10 and 7 for the genera of *Alphacoronavirus* (α–CoV), *Betacoronavirus* (β–CoV), *Deltacoronavius* (δ–CoV) and *Gammacoronaviru*s (γ–CoV), correspondingly, as shown in Table I.

TABLE I. Accession Codes for Prototype Variants of Coronaviruses in the GenBank

Genus	**GenBank ID**
α–CoV	NC_022103, NC_028814, NC_018871, NC_002645, NC_032730, NC_023760, KX512809, KX512810, EU420138, NC_010438, NC_28811, NC_028833, KT323979, NC_009657, NC_009988, AY567487, KY073745, KP981644, FJ938051, AY994055, KR270796, NC_038861
β–CoV	KF294357, BCU00735, KX432213, EF446615, AY391777, NC_017083, MF083115, NC_026011, AC_000192, KF294371, NC_012936, NC_006577, NC_025217, KF917527, JX869059, MG596803, MK679660, NC_009019, NC_009020, NC_030886, NC_009021, MG772933, MG772934, AY278489, FJ588686, NC_045512, MT121216, MN996532
δ–CoV	NC_016995, JQ065042, KJ569769, NC_016992, NC_016991, FJ376620, NC_011550, NC_016993, NC_016994, NC_016996
γ–CoV	EU111742, KF793826, KF696629, GQ504724, NC_010800, AY641576, MK423877

Besides the genomes of prototype variants of genus, the genomes of four coronavirus genera from the GenBank of release 237 (https://ftp.ncbi.nlm.nih.gov/genbank/) were used. Final quantities of all analyzed the genome sequences together with the genomes of prototype variants of each genus constitute: 924, 1954, 19 and 345 for the genera of α–CoV, β–CoV, δ–CoV and γ–CoV, correspondingly.

Let us introduce quantitative characteristics needed for the method proposed and for determinating which coronavirus genus the analyzed genome sequences belong to. The mean length over all genes analyzed in the work (over their coding regions – CDSs) for each coronavirus genus is shown in Table II in nucleotides (nt).

TABLE II. Mean Length of the Key Fragments from the Coronavirus Genomes

Genus	**ORF1ab, nt**	**S-gene, nt**	**N-gene, nt**
α–CoV	$L_{ORF}^{\alpha} = 20284$	$L_{S}^{\alpha} = 4105$	$L_{N}^{\alpha} = 1270$
β–CoV	$L_{ORF}^{\beta} = 21275$	$L_{S}^{\beta} = 3926$	$L_{N}^{\beta} = 1267$
δ–CoV	$L_{ORF}^{\delta} = 18794$	$L_{S}^{\delta} = 3590$	$L_{N}^{\delta} = 1044$
γ–CoV	$L_{ORF}^{\gamma} = 19861$	$L_{S}^{\gamma} = 3591$	$L_{N}^{\gamma} = 1229$

To calculate a mean distribution of codon frequency in every gene (ORF1ab, S- and N-gene), the coronavirus genomes of known prototype variants were used which accession codes in the GenBank are shown in Table I.

For genus x–CoV, where $x \in \{\alpha,\beta,\delta,\gamma\}$ is symbol denominating coronavirus genus, the quantitative characteristics are written as following:

M^x – is a quantity of prototype variants of x–CoV genus;

n – is a number of certain prototype variant;

$\boldsymbol{p}_{\mathrm{Y}}^{x}(n)$ – is a distribution of codon frequency in CDS of an Y gene, where $\mathrm{Y} \in \{\mathrm{ORF1ab, S, N}\}$ is symbol of gene considered.

While considering the Y gene, in frame of the denominations pointed above, for frequency codon distribution $\boldsymbol{P}_{\mathrm{Y}}^{x}$ that is averaged over the prototypes of genus x–CoV a formula is used:

$$\boldsymbol{P}_{\mathrm{Y}}^{x} = \frac{1}{M^x} \sum_{n=1}^{M^x} \boldsymbol{p}_{\mathrm{Y}}^{x}(n) . \qquad (1)$$

The results of numerical experiences showed that determining of coronavirus genus is improved significantly, if the lowest codon frequencies are excluded from the $\boldsymbol{P}_{\mathrm{Y}}^{x}$ distribution. It appears that for all genes three stop-codons (TERM) have such the frequencies. Besides, for ORF1ab together with S-gene the two (cga, cgg) of six synonymous arginine (ARG) codons have such the lowest frequencies, but for N-gene the frequencies of two cysteine (CYS) codons (tgt, tgc) are the lowest ones. These facts have been taken into account in calculating the averaged frequency codon distributions by (1). Therefore, format of averaged distribution $\boldsymbol{P}_{\mathrm{Y}}^{x}$ is written as $\boldsymbol{P}_{\mathrm{Y}}^{x} = (P_{\mathrm{Y}1}^{x}, P_{\mathrm{Y}2}^{x}, \ldots, P_{\mathrm{Y}59}^{x})$.

If coronavirus genome of unknown genus is under consideration, then for its distribution of codon frequency (after elimination of the codons pointed above) in Y gene a designation $\boldsymbol{p}_{\mathrm{Y}}$ is used. Format of the distribution $\boldsymbol{p}_{\mathrm{Y}}$ is written as $\boldsymbol{p}_{\mathrm{Y}} = (p_{\mathrm{Y}1}, p_{\mathrm{Y}2}, \ldots, p_{\mathrm{Y}59})$. With such the designation for the $\boldsymbol{p}_{\mathrm{Y}}$ distribution of codon frequency in analyzed single genome of prototype variant or in coronavirus genome of unknown genus a deviation $D(\boldsymbol{P}_{\mathrm{Y}}^{x}, \boldsymbol{p}_{\mathrm{Y}})$ from the $\boldsymbol{P}_{\mathrm{Y}}^{x}$ average distribution of codon frequency in Y gene of x–CoV genus is calculated by formulae:

$$D(\boldsymbol{P}_{\mathrm{Y}}^{x}, \boldsymbol{p}_{\mathrm{Y}}) = \frac{1}{7} \sum_{i=1}^{59} \frac{|P_{Yi}^{x} - p_{Yi}|}{P_{Yi}^{x}} . \qquad (2)$$

Among the deviations $D(\boldsymbol{P}_{\mathrm{Y}}^{x}, \boldsymbol{p}_{\mathrm{Y}})$, where $x \in \{\alpha,\beta,\delta,\gamma\}$, the minimal one is chosen, that points at genus of coronavirus analyzed according to the Y gene.

III. Results and Discussion

A. Identification of coronavirus genus for prototype variants

Table III shows the results of applying of the approach proposed to coronavirus genus identification which is based on learning sample of prototype variants.

As it follows from the Table III, applying of the method proposed to determinate coronavirus genus in the genes of learning sample showed the result in frame of allowable statistical error. On the average, reliability over all three key fragments constituted 93 %.

TABLE III. Results of Correct Genus Identification for Prototype Variants of Coronaviruses

Key Fragment	**α–CoV,** Totally 22 variants	**β–CoV,** Totally 28 variants	**δ–CoV,** Totally 10 variants	**γ–CoV,** Totally 7 variants
ORF1ab	20	25	9	7
S	22	23	9	7
N	19	28	10	7

B. Identification of genus amond the coronavirus genomes from the GenBank

On the strength of reasonably high reliability in determinating coronavirus genus that was achieved at learning sample of prototype variants, let us apply statistic (2) for determining genus of viral genomes from the GenBank, and in using the same averaged codon frequencies (1) that have been obtained for the prototypes of each coronavirus genus. Results of such genus identification for the coronaviruses from the GenBank database including the results for genus of prototype variants are shown in Table IV. According to Table IV reliability of identification, basing on ORF1ab, constitutes 98 %, on the other hand reliability of identification, basing on S- and N- genes, achieves 93 % only.

TABLE IV. Results of Correct Genus Identification for All Analyzed Genomes of Coronaviruses

Key Fragment	**α–CoV,** Totally 924 genomes	**β–CoV,** Totally 1954 genomes	**δ–CoV,** Totally 19 genomes	**γ–CoV,** Totally 345 genomes
ORF1ab	910	1906	16	343
S	898	1898	17	192
N	743	1924	19	343

Significant error for S- and N- genes is due to the following reasons.

Quantity of the whole sample of *Gammacoronavirus* genomes analyzed in the work from the GenBank is equal to 345 sequences (see Table IV). However, in genus identification, basing on spike-protein S-gene, in 143 cases of the γ–CoV genes the genera of α–CoV (15.4%) and β–CoV (84.6%) were identified by statistic (2).

Analogous phenomenon was observed when α–CoV genus was recognizing on the base of N-gene of nucleocapside protein. Though for the genera β–CoV, δ–CoV and γ–CoV identification, basing on the N-gene, reliability is, practically, of 100 %, in recognizing α–CoV genus there are 131 events when statistics (2) point at β–CoV genus.

So, maximal reliability in determining genus of coronavirus is achieved, if ORF1ab is used.

Ambiguous results obtained in the work in determining genus of coronavirus on the base of S-gene and N-gene may be explained by homologous recombinations, that occur among the coronavirus genomes of different genera and promote mosaicism of the genome structure [8, 9].

IV. Conclusion

Method for determining coronavirus genus on the base of statistic, that efficiency was earlier shown in recognizing species of the flaviviruses, is proposed in the work. The statistic uses the codon frequency distributions in the genes of coronaviruses. Three genes were considered as key genome fragments. The first one is ORF1ab, encoding a number of viral non-structural proteins. The rest two are S- and N- genes which encode structural proteins (spike and nucleocapsid proteins) of coronavirus virion. The method proposed was developed with the help of learning sample of the coronavirus genomes being the prototype variants for the four genera.

Implication of the proposed statistics for determining coronavirus genus on the base of each gene (key genome fragment) mentioned above showed acceptable reliability. The best result of reliability (98%) was achieved at ORF1ab, encoding non-structural proteins.

From the other hand, comparing identification of coronavirus genus on the base of the three genes (ORF1ab, S и N), a phenomenon of mosaicism in the coronavirus genome was revealed, that is due to homologous recombination of the genes. So, for example, among analyzed the 345 γ–CoV genomes α–CoV genus was "recognized" 22 times and β–CoV genus was "determined" in 127 cases. As well as among all considered the 924 genomes of α–CoV genus in 170 events β–CoV genus was "recognized" basing on N-gene of nucleocapsid protein.

Phenomenon of mosaicism may be supposed as a reflection of plasticity of the coronavirus genome and some kind of its readiness not only to expansion of novel life areas but searching for new hosts also.

The method proposed, particularly, may be used both in metagenomic investigations of microbial content in nature environment and in determining taxonomic affiliation of the viruses obtained.

References

[1] "Virus taxonomy. Classification and nomenclature of viruses. Ninth report of the International Committee on Taxonomy of Viruses." A. M. Q. King, M. J. Adams, E. B. Carstens, E. J. Lefkowitz, Eds. Elsevier Academic Press, 2011.

[2] J. Ziebuhr, R. S. Baric, S. Baker, R. J. de Groot, C. Drosten, A. Gulyaeva, B. L. Haagmans, B. W. Neuman, S. Perlman, L. L. M. Poon, I. Sola, A. E. Gorbalenya, "Reorganization of the family *Coronaviridae* into two families, *Coronaviridae* (including the current subfamily Coronavirinae and the new subfamily *Letovirinae*) and the new family *Tobaniviridae* (accommodating the current subfamily *Torovirinae* and three other subfamilies), revision of the genus rank structure and introduction of a new subgenus rank," Proposal 2017.013S (08.08.2018) for International Committee on Taxonomy of Viruses.

[3] K. Bukhari, G. Mulley, A. A. Gulyaeva, L. Zhao, G. Shu, J. Jiang, B. W. Neuman, "Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family *Abyssoviridae*, and from a sister group to the *Coronavirinae*, the proposed genus *Alphaletovirus*," Virology, vol. 524, pp. 160–171, 2018.

[4] D. K. Lvov, S. V. Alkhovsky, "Source of the COVID-19 pandemic: ecology and genetics of coronaviruses (*Betacoronavirus*: *Coronaviridae*) SARS-CoV, SARS-CoV-2 (subgenus *Sarbecovirus*), and MERS-CoV (subgenus *Merbecovirus*)," Problems of virology, vol. 65 (2), pp. 62–70, 2020. (In Russ.)

[5] B. W. Neuman, B. D. Adair, C. Yoshioka, J. D. Quispe, G. O. P. Kuhn, R. A. Milligan, M. Yeager, M. J. Buchmeier, "Supramolecular architecture of severe acute respiratory syndrome coronavirus revealed by electron cryomicroscopy," J. Virol., vol. 80 (16), pp. 7918–7928, 2006.

[6] Virology: principles and applications. J. Carter, V. Saunders, Eds. Chichester, England: John Wiley & Sons Ltd, 2007.

[7] M. Yu. Shchelkanov, A. Yu. Popova, V. G. Dedkov, V. G. Akimkin, V. V. Maleev, "History of investigation and current classification of coronaviruses (*Nidovirales*: *Coronaviridae*)," Russian Journal of Infection and Immunity = Infektsiya i immunitet, vol. 10 (2), pp. 221–246, 2020.

[8] Y. Chen, Q. Liu, D.Guo, "Emerging coronaviruses: Genome structure, replication, and pathogenesis," J. Med. Virol., vol. 92, pp. 418–423, 2020.

[9] M. M. C. Lai, "Recombination in large RNA viruses: Coronaviruses," Seminars in Virology, vol. 7 (6), pp. 381–388, 1996.

[10] Y.Tao, M. Shi, C. Chommanard, K. Queen, J. Zhang, W. Markotter, I. V. Kuzmin, E. C. Holmes, S. Tong, "Surveillance of bat coronaviruses in Kenya identifies relatives of human coronaviruses NL63 and 229E and their recombination history," Journal of Virology, vol. 91 (5), pp. e01953–16, 2017.

[11] Y. Ma, L. Wu, Shaw N., Y. Gao, J. Wang, Y. Sun, Z. Lou, L. Yan, R. Zhang, Z. Rao, "Structural basis and functional analysis of the SARS coronavirus nspl4-nspl0 complex," PNAS, vol. 112 (30), pp. 9436–9441, 2015.

[12] D. Cavanagh, K. Mawditt, A. Adzharet et al. "Does IBV change slowly despite the capacity of the spike protein to vary greatly?" Adv. Exp. Med. Biol., vol. 440, pp. 729–734, 1998.

[13] Y. Chen, Q. Liu, D. Guo, "Emerging coronaviruses: Genome structure, replication, and pathogenesis," J. Med. Virol., vol. 92. pp. 418–423, 2020.

[14] M. B. Chaley, Zh. S. Tyulko, V. A. Kutyrkin, "Flavivirus Species Recognition Based On the Polyprotein Coding Sequences," Math. Biol. Bioinf., vol. 14(2), pp. 533–542, 2019. (In Russ.)