Optimization of a coronavirus genus recognition procedure based on the n-gene of prototypic strains

Maria Chaley[1,*], and *Vladimir* Kutyrkin[2]

[1]IMPB RAS – Branch of Keldysh Institute of Applied Mathematics RAS, Pushchino, Russia
[2]Moscow State Technical University n.a. N.E. Bauman, Moscow, Russia

Abstract. The article offers a solution to the problem of fast and efficient recognition of the coronavirus genus. For this purpose, the authors apply a virus genome targeting method based on the use of a sufficiently short and conserved N-gene of the nucleocapsid protein. Comparison of the codon frequency distributions in the N-gene of the analyzed genome and a set of 67 prototypical strains corresponding to the coronavirus subgenus allows us to recognize the genus of the coronavirus. This paper proposes optimization of the genus recognition of coronavirus by eliminating a significant number of codons from the 64 codons of the genetic code (26 in one case and 57 in the other). The authors achieved 100% genus recognition efficiency in a sample of 2,051 coronavirus genomes from the GenBank database with annotated subgenus in the optimized procedure. The authors also achieved 99% confidence when using the optimized coronavirus genus recognition procedure in a total sample of 3,242 genomes.

1 Introduction

Coronaviruses (Coronaviridae) are a large family of RNA-containing viruses capable of infecting humans and animals. The COVID-19 pandemic caused by SARS-CoV-2 has brought renewed attention to coronaviruses and their evolutionary potential [1]. High mutation and recombination rates contribute to the genotypic and phenotypic variability of coronaviruses [2]. Many scientists have studied the transmission of the virus from the main host population to other animal species and humans at the molecular genetic level [3]. Modern methods of analyzing viral genomes can create a system for monitoring gene pools of viral populations in unique ecosystems [4].

A reasonably fast and cost-effective targeted sequencing approach helps in tracking mutant strains of SARS-CoV-2 and other coronavirus species [5]. Here, only the sequence of the most variable fragment of the S-protein gene responsible for binding to cellular receptors during infection was determined. We can also propose a targeting approach for the rapid identification of coronavirus species, genus, and subgenus based on separately isolated fragments of the viral genome.

* Corresponding author: maramaria@yandex.ru

Work [6] examined various approaches to the recognition of the coronavirus genus (CoVs) based on the genomes of prototypical coronavirus strains. The authors characterized the coronavirus genome by the distribution of codon frequencies of its individual structural and nonstructural genes: the M gene for membrane protein, the S gene for spike protein, the N gene for nucleocapsid protein, and the ORF1ab gene encoding several nonstructural proteins. Scientists divided each of the four genera (α-CoV, β-CoV, δ-CoV, γ-CoV) into subgenera that also have several prototypic strains. The coronavirus genus was characterized by an analytical, i.e. averaged, distribution of codon frequencies of all its prototypic strains. In the variant approach [6] to genus recognition, each variant (regarded as a target) relied on different combinations of structural and non-structural genes. The variant based on the N-gene of the nucleocapsid protein showed the best result in recognizing the genera β-CoV, δ-CoV, and γ-CoV. Averaging the codon frequency distributions of prototypic strains of a significant number (12) subgenera in α-CoV caused a low recognition efficiency of this genus.

This paper was aimed at fast and reliable identification of the genus from a fragment (target) of the coronavirus genome. The authors chose the N-gene among the previously identified targets as the most conservative and significantly shorter (~1200 nucleotides) compared to other genes. The N-gene encoded nucleocapsid protein has functions related to viral pathogenesis, transcription and replication. Scientists often use the N-gene for molecular diagnostics of CoVs [7-9].

We moved from recognizing the genus of coronavirus based on analytical (averaged) codon frequency distributions [6] to recognition based on individual frequency distributions in the N-gene of prototypic subgenus strains. Determining the subgenus of a coronavirus automatically determines its genus. We investigated the dependence of recognition efficiency on the use of different codon groups in the N-gene. The groups were formed according to selected levels of codon frequencies averaged over the N-gene of 67 prototypic strains of all four coronavirus genera. We identified the two most efficient groups of codons, including 38 codons in one case and 7 codons in the other. Reducing the number of codons used optimized the recognition procedure. This optimization improved the efficiency of coronavirus genus recognition compared to recognition based on the analytical mean distributions of the coronavirus genus of the paper [6].

2 Materials

We used coronavirus genomes from four genera from the GenBank database: 3242 pieces, of which for 2051 genomes a subgenus was specified beside the genus. We used the codon distributions in the N genes of the prototypic strains from GenBank in the genus recognition procedure. Table 1 contains the access codes of the prototypical coronavirus strains in GenBank.

Table 1. GenBank access codes for prototypic coronavirus strains. Subgenus designations and indices of prototypic strains used in this work.

Genus	Subgenus	GenBank ID	Variant index k
α-CoV	*Colacovirus* (α01)	NC_022103	1
	Decacovirus (α02)	NC_028814	1
		NC_018871	2
	Duvinacovirus (α03)	NC_002645	1
	Luchacovirus (α04)	NC_032730	1
	Minacovirus (α05)	NC_023760	1
		KX512809	2
		KX512810	3

Continuation of Table 1.

β-CoV	*Minunacovirus* (α06)	EU420138	1
		NC_010438	2
	Miotacovirus (α07)	NC_028811	1
	Nyctacovirus (α08)	NC_028833	1
	Pedacovirus (α09)	KT323979	1
		NC_009657	2
	Rhinacovirus (α10)	NC_009988	1
	Setracovirus (α11)	AY567487	1
		KY073745	2
	Tegacovirus (α12)	NC_038861	1
		KP981644	2
		FJ938051	3
		AY994055	4
		KR270796	5
	Embecovirus (β01)	KF294357	1
		BCU00735	2
		KX432213	3
		EF446615	4
		AY391777	5
		NC_017083	6
		MF083115	7
		NC_026011	8
		AC_000192	9
		KF294371	10
		NC_012936	11
		NC_006577	12
	Hibecovirus (β02)	NC_025217	1
	Merbecovirus (β03)	KF917527	1
		JX869059	2
		MG596803	3
		MK679660	4
		NC_009019	5
		NC_009020	6
	Nobecovirus (β04)	NC_030886	1
		NC_009021	2
	Sarbecovirus (β05)	MG772933	1
		MG772934	2
		AY278489	3
		FJ588686	4
		NC_045512	5
		MT121216	6
		MN996532	7
δ-CoV	*Andecovirus* (δ01)	NC_016995	1
	Buldecovirus (δ02)	JQ065042	1
		KJ569769	2
		NC_016992	3
		NC_016991	4
		FJ376620	5
		NC_011550	6
		NC_016993	7
	Herdecovirus (δ03)	NC_016994	1
	Moordecovirus (δ04)	NC_016996	1

Continuation of Table 1.

γ-CoV	*Cegacovirus* (γ01)	EU111742	1
		KF793826	2
	Igacovirus (γ02)	KF696629	1
		GQ504724	2
		NC_010800	3
		AY641576	4
		MK423877	5

Next to the name of the subgenus in parentheses is its designation in this paper (Table 1). We considered 67 prototypic strains, each of which is indexed in its subgenus. Work [10] provides details of species with prototypic coronavirus strains.

3 Methods

With the high mutation rate of viruses, it would be interesting to assess which N-gene codons determine whether a coronavirus belongs to a subgenus. This raises the question of a reduction in the number of codons in the N-gene of a coronavirus to recognize its subgenus. We selected the N-gene codons relative to their average frequency of occurrence in all subgenus of the prototype strains (see Table 1). Figure 1 shows a graph of the average frequency of codons in the N-gene of prototypic coronavirus strains. Thirty-eight codons occur with frequencies above the threshold shown by the dashed line. Seven codons occur with frequencies above the threshold shown by the dotted line.

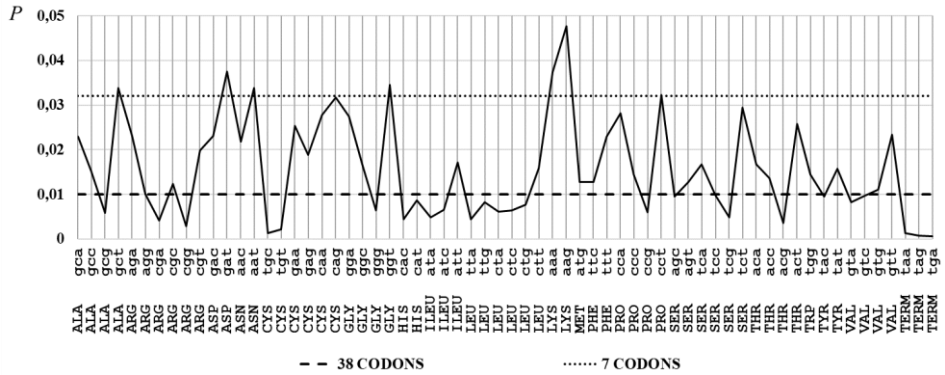

Fig. 1. Average frequency *P* of occurrence of codons in the N-gene of prototypic coronavirus strains.

Selecting a codon frequency threshold means that only those codons whose frequency exceeds the threshold are suitable for recognition. Figure 1 shows two such thresholds and the number of codons whose frequencies exceed each of them with different horizontal lines. Together with the codons, it shows the amino acids they encode along the horizontal axis.

Let the chosen threshold leave the set of $C_r = \{c_1, c_2, ..., c_r\}$ N-gene codons, where $|C_r| = r$ – the number of codons in this population. ρ_i – the frequency of occurrence of a codon c_i in the N-gene of the coronavirus in question $i = \overline{1,r}$. The string $\boldsymbol{\rho} = (\rho_1, \rho_2, ..., \rho_r)$ characterizes the coronavirus genome under analysis.

$S \in \{\alpha, \beta, \delta, \gamma\}$ is the symbol for the genus of the coronavirus, j is the two-digit subgenus number in the genus, and k is the index of the strain in this subgenus according to Table 1. For example, $\alpha 12, 4$ is the notation of the prototype strain AY994055 with an index

$k = 4$ of the subgenus *Tegacovirus* from α-CoV. Then $\rho_i^{Sj,k}$ – the frequency of occurrence of codon c_i in the N-gene of the prototypic strain Sj,k, where Sj is the subgenus of the strain (and k is the index of the strain in that subgenus). Thus, the prototypic strain Sj,k is characterized by the string $\boldsymbol{\rho}^{Sj,k} = (\rho_1^{Sj,k}, \rho_2^{Sj,k}, ..., \rho_r^{Sj,k})$.

The number $\Delta^{Sj,k}(\boldsymbol{\rho})$ calculated using the formula (1) defines the deviation of the coronavirus genome with a frequency distribution of $\boldsymbol{\rho}$ from the prototypical strain Sj,k.

$$\Delta^{Sj,k}(\boldsymbol{\rho}) = \frac{1}{7}\sum_{i=1}^{r}\frac{|\rho_i - \rho_i^{Sj,k}|}{\rho_i^{Sj,k}}, \quad (1)$$

where the coefficient $\frac{1}{7}$ is chosen for ease of presentation of results.

Having got the distribution deviations (1) from all the distributions of the prototypical strains, we choose the strain Sj,k with the minimum deviation. We hypothesize that the coronavirus genome in question belongs to subgenus Sj and thus to genus S (see Table 1).

4 Results and discussion

We denote $M(r)$ the number of errors in the coronavirus's recognition subgenus using a set of codons C_r and formula (1) for the 2051 coronavirus genome annotated by the subgenus. Figure 2 shows the dependence of $M(r)$ the number of errors on the number of r codons used.

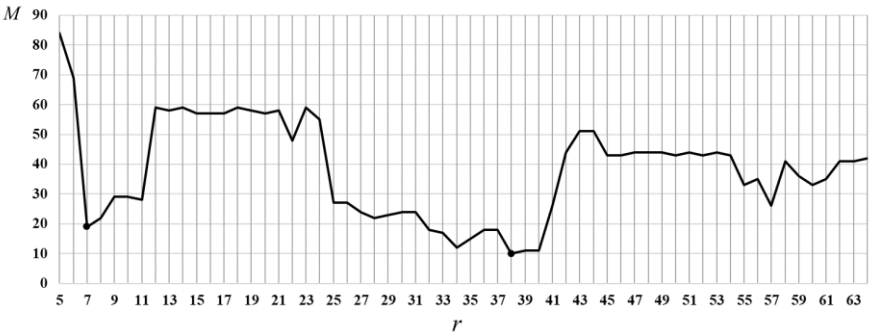

Fig. 2. Plot of the number of subgenus recognition errors in 2051 N-genes of coronaviruses as a function of the size r of the set of the genetic code codons used for recognition.

Analysis of the graph in Figure 2 reveals two particular minima on sets of 38 and seven codons. The set of 38 codons achieves the minimum number of errors ($M = 10$). The analysis of Figure 2 reveals a sharp minimum ($M = 19$) from a set of seven codons. Figure 3 shows the codons that are part of the aggregates C_{38} and C_7.

	C_{38}	C_7		C_{38}	C_7		C_{38}	C_7		C_{38}	C_7
LYS aaa	1	1	THR aca	1	0	ALA gct	1	1	LEU ctt	1	0
LYS aag	1	1	THR acc	1	0	ALA gca	1	0	LEU cta	0	0
ASN aat	1	1	THR act	1	0	ALA gcc	1	0	LEU ctc	0	0
ASN aac	1	0	THR acg	0	0	ALA gcg	0	0	LEU ctg	0	0
ARG cgt	1	0	MET atg	1	0	VAL gtt	1	0	LEU tta	0	0
ARG cgc	1	0	ILEU att	1	0	VAL gtg	1	0	LEU ttg	0	0
ARG cga	0	0	ILEU ata	0	0	VAL gta	0	0	PHE ttt	1	0
ARG cgg	0	0	ILEU atc	0	0	VAL gtc	0	0	PHE ttc	1	0
ARG aga	1	0	ASP gat	1	1	CYS caa	1	0	TYR tat	1	0
ARG agg	1	0	ASP gac	1	0	CYS cag	1	0	TYR tac	0	0
SER agt	1	0	CYS gaa	1	0	HIS cac	0	0	TERM taa	0	0
SER agc	0	0	CYS gag	1	0	HIS cat	0	0	TERM tag	0	0
SER tca	1	0	GLY ggt	1	1	PRO cct	1	1	TERM tga	0	0
SER tct	1	0	GLY gga	1	0	PRO cca	1	0	TRP tgg	1	0
SER tcc	0	0	GLY ggc	1	0	PRO ccc	1	0	CYS tgc	0	0
SER tcg	0	0	GLY ggg	0	0	PRO ccg	0	0	CYS tgt	0	0

Fig. 3. The C_{38} and C_7 aggregates of the genetic code codons with the lowest number of subgenus recognition errors by the N-gene of coronavirus. The number 1 shows codon entry into the aggregate, 0 shows no codon in the aggregate. To the left of the codon is the three-letter designation of the amino acid encoded. Stop codons are designated as TERM.

Note that there are no genus recognition errors out of 10 for subgenus recognition based on 38 codons. When subgenus recognition is based on 7 codons out of 19 errors, there are 14 errors in genus recognition of coronavirus.

If we consider a combined sample of genus and subgenus annotated (2051 genomes) and genus-only annotated (1191 genomes) coronavirus genomes, recognition of coronavirus genus based on 38 codons will lead to 10 errors and recognition based on 7 codons to 34 errors.

The proposed methods for subgenus and genus recognition of coronaviruses yield at least 99% confidence.

5 Conclusions

When using the above two approaches for genus recognition based on N genes, out of 2051 genomes annotated for the subgenus in the first case (with averaging), we noted 43 errors (of which 19 were incorrectly determined by the genus coronavirus) and in the second (without averaging); we noted 36 errors, of which only three were incorrectly determined by the genus. This result shows that the use of analytical (averaged) codon frequency distributions of prototypic strains is less effective for recognizing both the subgenus and genus of coronaviruses.

References

1. D. K. Lvov, S. V. Alkhovsky, L. V. Kolobukhina, E. I. Burtseva, Problems of Virology **65**, 6 (2020)
2. E. Salem, V. Dhanasekaran, H. Cassard, B. Hause, S. Maman, G. Meyer, et al., Viruses **12**, 534 (2020)
3. E. Corrales-Aguilar, M. Schwemmle Eds., *Bats and Viruses: Current Research and Future Trends* (Academic Press, Caister, 2020)
4. D. K. Lvov, S. V. Borisevich, S. V. Alkhovsky, E. I. Burtseva, Infectious Diseases: News, Opinions, Training **8**, 96 (2019)
5. N. I. Borisova, I. A. Kotov, A. A. Kolesnikov, V. V. Kaptelova, A. S. Speranskaya, L. Yu. Kondrasheva, et al., Problems of Virology **66**, 269 (2021)
6. M. B. Chaley, V. A. Kutyrkin, Math. Biol. Bioinf. **17**, 10 (2022)

7. L. J. Saif, Vet. Clin. North Am. Food Anim. Pract. **26**, 349 (2010)
8. A. N. Vlasova, L. J. Saif, Front. Vet. Sci. **8**, 643220 (2021)
9. A. G. Glotov, A. V. Nefedchenko, A. G. Yuzhakov, S. V. Koteneva, T. I. Glotova, A. K. Komina, N. Yu. Krasnikov, Problems of Virology **67**, 465 (2022)
10. M. Yu. Shchelkanov, A. Yu. Popova, V. G. Dedkov, V. G. Akimkin, V. V. Maleev, Russian Journal of Infection and Immunity **10**, 221 (2020)