

Variant approach to genus identification of coronavirus genome

Chaley M.^{1*}, Kutyrkin V.²

¹ *Institute of Mathematical Problems of Biology, RAS – Branch of Keldysh IAM RAS, Pushchino, Russia*

² *Moscow State Technical University n.a. N.E. Bauman, Moscow, Russia*

* *chaley@phystech.edu*

Key words: coronavirus genome, ORF1ab, S-gene, M-gene, N-gene, statistical analysis, variant approach, coronavirus genus identification

A new variant approach to identification of coronavirus genus basing on modified statistic which earlier showed efficiency for flavivirus species recognition is proposed. This statistic is founded on codon frequency distribution in the genes of viral genome. ORF1ab encoding non-structural proteins and the genes of coronavirus structural proteins (S, M and N) are under consideration. The variant approach uses both combinations of the genes and singular genes that are named variants. Six variants are considered in the work. In weighting length of each gene in a variant, joint distribution of codon frequency is calculated. Genus of coronavirus is determined in the result of analysis of all the variants considered. The approach was developed with the help of learning samples of prototype variants for the four genus of the coronaviruses: α -CoV, β -CoV, δ -CoV and γ -CoV. Results of genus identification for certain genome is represented by variant string $\mathbf{v} = (v_1, v_2, v_3, v_4, v_5, v_6)$, where $v_i \in \{\alpha, \beta, \delta, \gamma\}$ is the result of coronavirus determining on the i th ($i = \overline{1, 6}$) combination of the genes. If at least four out of the six components point at the same genus of coronavirus, then such the genus is recognized as the result of identification. According to this rule reliability of genus identification for more than the 3000 of coronavirus genomes from the GenBank was of 97 % [1]. In the majority of events the components of variant string are represented by the same character. However, the cases are frequent enough when a few components in the string are different from the others. So, the variant approach reveals a mosaicism in determining genus of the coronavirus genome. It seems likely that phenomenon of mosaicism is due to high genetic variability and propensity to homologous recombinations in interpopulation interactions of the coronaviruses and their natural reservoirs.

References

1. Chaley M., Kutyrkin V. Coronavirus genus recognition based on prototype virus variants. *Math Biol Bioinf.* 2022;17(1):10-27. (in Russian)