doi 10.18699/bgrs2024-12.3-21

Mutation in of palindromic sequences of coronavirus genome

Mitić N.S.[1]*, Kapunac S.[1], Beljanski M.V.[1], Dergilev A.I.[2], Orlov Y.L.[2, 3]**

[1] *University of Belgrade, Belgrade, Serbia*

[2] *Novosibirsk State University, Novosibirsk, Russia*

[3] *Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow, Russia*

** nenad.mitic@matf.bg.ac.rs, ** orlov@bionet.nsc.ru*

Key words: computer genomics, microbiology, coronavirus, mathematical modelling, mutation rate

Motivation and Aim: The RNA genome of SARS-CoV-2 was shown to be organized into structural and functional blocks of RNA information that are demarcated by short RNA breakpoint sequences that promote recombination at specific non-random locations within the viral genome consisting of short repetitive sequences, namely palindromes. Palindromic sequences are involved in the formation of RNA secondary structures. They can be locations recognized by RNA-binding proteins as well as places of RNA recombination [1]. We analyzed SARS-COV-2 genomes with particular attention to mutations within palindromic sequences.

Methods and Algorithms: A dataset of 423.425 complete isolate nucleotide sequences was extracted from https://www.ncbi.nlm.nih.gov/sars-cov-2. After the cleanup process, 347.962 isolates with 123.667 unique (related nucleotide sequences) with 226.624 corresponding unique protein (nucleotide) coding sequences and 141.926 unique protein (AA) sequences remain. The consistency of the two sequences was checked using the standard genetic code table (transl_table 1). Each sequence was annotated with a World Health Organization (WHO) SARS-CoV-2 annotation. Each nucleotide sequence was individually aligned to the reference sequence SARS-COV -2 (NC_045512.2) using the MAFFT alignment program [2]. The StatRepeats program [3] was used to determine all palindromes with a minimum length of 8. A total of 801.935.394 palindromes were determined. Among them, 785.854.841 repeats were identical with their pair in reference sequence NC_045512.2. Other (16.080.553) palindromes have some mutations related to reference sequence. The analysis of the number of palindrome occurrences was performed in 5 time intervals of 4 months (pandemic time from 31.12.2019 to 25.08.2021). The average number of palindromes per isolate shows a constant increase, respectively by intervals: 1.92, 3.51, 9.31, 14.84, and 20.66.

Results: The pipeline for the analysis of viral genome in relation to mutation rate is presented. We analyzed mutations in all 12 types of ORFs present in the set of extracted sequences (ORF1a polyprotein, ORF1ab polyprotein, surface glycoprotein, ORF3a protein, envelope protein, membrane glycoprotein, ORF6 protein, ORF7a protein, ORF7b protein, ORF8 protein, nucleocapsid phosphoprotein, ORF10 protein). Among them, normalized on average protein length, after ORF1a and ORF1ab,the surface glycoprotein (S-protein) has the highest number of repeats, on average 4.65 palindromes with a length of > =8. The highest number of palindromes is located around positions 22.000 (left part) and 24.300 (right part), counting the positions with respect to the beginning of the isolates. For the total number of mutations, almost 78% resulted in amino-acid changes in the corresponding proteins.

Conclusion: Available tools and the problems of mutation rate dependence on sequence context will be discussed. We discussed this work at VII Congress on Biophysics [4]. The complexity measures and the entropy estimates of genome segments help to reveal mutated profile of the sequences [5]. We plan to perform a detailed analysis of mutations of palindromic sequences according to the SARS-CoV-2 WHO variant classification and also their influence on amino-acid changes and occurring RNA secondary structures or locations recognized by RNA binding proteins.

Acknowledgments: The authors are grateful to the colleagues from Faculty of Mathematics, Belgrade University for science discussion.

References

1. Gallaher W.R. A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2. *Arch Virol*. 2020;165:2341-2348. doi 10.1007/s00705-020-04750-z
2. Katoh K., Rozewicki J., Yamada K.D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform.* 2019;20(4):1160-1166. doi 10.1093/bib/bbx108
3. Jelovic A., Mitic N., Eshafah S., Beljanski M. Finding Statistically Significant Repeats in Nucleic Acids and Proteins. *J Comput Biol.* 2018;25:375-387. doi 10.1089/cmb.2017.0046
4. Anashkina A.A., Rubin A.B., Gudimchuk N.B., Vanin A.F., Tsygankov A.A., Orlov Y.L. VII Congress of Russian Biophysicists-2023. *Biophys Rev*. 2023;15(5):801-805. doi 10.1007/s12551-023-01164-4
5. Orlov Y.L., Orlova N.G. Bioinformatics tools for the sequence complexity estimates. *Biophys Rev*. 2023;15(5):1367-1378. doi 10.1007/s12551-023-01140-y