

Approaches for the coronavirus genomes annotation

Korneenko E.V.^{1, 2*}, Chudinov I.K.^{1, 3}, Butenko I.O.¹, Speranskaya A.S.¹

¹ Scientific Research Institute for Systems Biology and Medicine, Federal Service on Consumer Rights Protection and Human Well-Being Surveillance, Moscow, Russia

² Saint Petersburg Pasteur Institute, Federal Service on Consumer Rights Protection and Human Well-Being Surveillance, St. Petersburg, Russia

³ Moscow Institute of Physics and Technology (National Research University), Moscow, Russia

* lenakorneenko0@gmail.com

Key words: viruses; coronaviruses; genome annotation

Motivation and Aim: The discovery of novel viruses especially coronaviruses (CoVs) has great importance in public health [1]. After the genome assembly each sequence should be annotated. Accurate gene prediction is a challenge that persists for decoding a newly sequenced genome [2]. To describe and visualize features of newly obtained viral genomes we applied the three step approach which included similarity-based prediction of genes/cds, protein domains prediction, output verification and manual correction.

Methods and Algorithms: Similarity-based prediction of genes/cds for assembled coronaviruses was performed using several programs such as Geneious Prime (<https://www.geneious.com>), VIGOR4 (a part of BV-BRC web service) [3] and Prokka, protein domains were predicted using the InterProScan tool against InterPro database [4]. Two alphacoronaviruses (Alpha-CoVs) belonging to *Pedacovirus* subgenus found in bats were annotated using Geneious Prime and Prokka [5]. Two betacoronaviruses (Beta-CoVs) belonging to the *Merbecovirus* subgenus found in hedgehogs were annotated using BV-BRC and Prokka. The output of annotators was manually checked, compared to the closest reference annotation using NCBI blast n, blast x, blast p and corrected.

Results: We performed gene/cds and protein domain prediction for 4 newly discovered complete genomes of coronaviruses: 2 Alpha-CoVs found in bats and 2 Beta-CoVs found in hedgehogs. We could predict 7 of 7 expected cds in *Pedacovirus* subgenus of Alpha-CoVs with Geneious prime, while Prokka found 8 cds. BV-BRC showed the presence of 10 of 11 mandatory cds for Beta-CoVs related to the *Merbecovirus* subgenus and Prokka showed the presence of 7 of 11 expected cds. Protein domain prediction using InterProScan successfully predicted typical Alpha-CoV domains in non-structural proteins. Protein domain annotation revealed that two Alpha-CoVs both contained large insertions of 130+ a.a. in the Nsp3 sequence. This region was labeled as “protein domain with unknown function”.

Conclusion: Annotation programs suitable for virus genomes (Prokka, BV-BRC and Geneious Prime) predicted most of expected cds in newly assembled CoVs as well as InterProScan was successful in annotation of protein domains in these sequences, but accurate gene and protein domain prediction required detailed knowledge of *Coronaviridae* genome structure, manual checking and correction. Thus, development of specialized algorithms for coronaviral genomes’ annotation that account for all known phenomena is a relevant task.

Funding: The study is supported by the state project “Development of algorithms to identify novel, unique DNA or RNA sequences in metagenomes and their phenotypic characterization in vitro” (No. 12203090069-4).

References

1. van Boheemen S., de Graaf M., Lauber C. et al. Genomic Characterization of a Newly Discovered Coronavirus Associated with Acute Respiratory Distress Syndrome in Humans. *mBio*. 2012;3(6):e00473-12. doi 10.1128/mBio.00473-12
2. Mills R., Rozanov M., Lomsadze A. et al. Improving Gene Annotation of Complete Viral Genomes. *Nucleic Acids Res*. 2003;31(23):7041-7055. doi 10.1093/nar/gkg878
3. Olson R.D., Assaf R., Brettin T. et al. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): A Resource Combining PATRIC, IRD and ViPR. *Nucleic Acids Res*. 2023;51(D1):D678-D689. doi 10.1093/nar/gkac1003
4. Paysan-Lafosse T., Blum M., Chuguransky S. et al. InterPro in 2022. *Nucleic Acids Res*. 2023;51(D1):D418-D427. doi 10.1093/nar/gkac993
5. Seemann T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics*. 2014;30(14):2068-2069. doi 10.1093/bioinformatics/btu153