The British Journal of Biology Studies
DOI: 10.32996/bjbs
Journal Homepage: www.al-kindipublisher.com/index.php/bjbs

Genome Sequence Analysis of Beta Coronavirus by Applying Bioinformatics Tools

Stuti[1] ✉ and Uma Kumari[2]
[1]P.G Student, Department of biotechnology final year (4th semester), Jiwaji University, Gwalior, M.P., India
[2]Bioinformatics Scientist Department of Bioinformatics, Rapture International Pvt Ltd, Noida, UP, India
✉ **Corresponding Author:** Stuti, **E-mail**: stutisanjit1604@gmail.com

ARTICLE INFORMATION

Received: 08 October 2021
Accepted: 25 October 2021
Published: 15 November 2021
DOI: 10.32996/bjbs.2021.1.1.4

KEYWORDS

Betacoronaviruses; Genomics; SARS-CoV-2; COVID-19, Coronavirus; Genome; Bioinformatics, COVID-19

ABSTRACT

The global pandemic crisis caused COVID-19 a disease with an alarming rate of human morbidity and case fatality produced by the currently emerging pathogen of SARS-CoV-2. The increase in the number of coronaviruses discovered and coronavirus genome sequencing has given us an idea to perform and utilize genomics and bioinformatics analysis on this particular family of viruses. These help us to understand the pathogenesis, animal origin, and mode of transmission of coronaviruses and have been used to tackle outbreaks caused by emerging, highly pathogenic, betacoronavirus strains, particularly emphasizing on SARS-CoV-2. SARS-CoV2 for comparing biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of RNA and/or DNA sequences. There have been identified more than 7,000 complete genome entries uploaded to Nucleotide/NCBI databases from the Coronaviridae family in the middle of 2002–2020, more than half of them are being SARS-CoV-2 and it is increasing in analogous to the expansion of sequencing technology. The BLASTP algorithm compares the protein database sequence with the query protein. FASTA produces a local alignment score in the compassion of the query sequence to every sequence in the database. Multiple sequence alignment residues provide comparative structure and functional analysis of biological sequences which often leads to fundamental biological insight into sequence-structure-function relationships of nucleotide or protein sequence families. Open reading frame indicates the protein-coding region in an RNA sequence with useful insight into genome structure and organization as well as the evolution of species. The contributions of bioinformatics for the planning and development of new drugs and the analysis of already known compounds support the search for safer and more effective treatments against SARS-CoV-2 infection.

1. Introduction

The serious rise in the number of coronaviruses originated and its genomes being sequenced have given us an unexampled opportunity to function genomics and bioinformatics analysis on this particular family of viruses. Since December 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has globally affected 195 countries. March 2020, the World Health Organization declared COVID-19 a pandemic. Then, COVID-19 has become overwhelmingly present in our lives, and researchers are working tirelessly to understand the virus. More than 17,000 papers by the query "COVID-19" indexed in PubMed/NCBI, nearly 200 of them, as of 29th May 2020, about the genome of severe acute respiratory syndrome-coronaVirus-2 (SARS-CoV-2) (Llanes et al. 2020) Coronaviruses were first isolated from chickens in 1937. There are now approximately 15 species in this family. Coronavirus particles are irregularly shaped, round about 60-220 nm in diameter, with an outer envelope bearing distinctive, ' club-shaped' peplomers (round about 20nm long x 10 nm at the wide distal end. This ' crown-like ' appearance (Latin, corona) gives the family its name. Coronaviruses are a group of associated RNA viruses that are enveloped, single-stranded, and positive-sense and a nucleocapsid of helical symmetry with a genome size of 29,903 bp. (Woo et al. 2010).

1.1 *Genomic Organization of SARS-CoV2*

SARS-CoV-2 is a single-stranded positive RNA virus of ~ 29.9 kB in size. The SARS-CoV-2 genome has 14 open reading frames (ORFs), which encodes for 27 different types of proteins. It has 5′ untranslated region (UTR), a replication complex (ORF1a and ORF1b), Spike (S) gene, Envelope € gene, Membrane (M) gene, Nucleocapsid (N) gene, 3′ UTR, several unidentified non-structural ORFs, and a poly (A) tail. ORF1a gene is located at the 5′UTR encodes for polyprotein pp1a, which contains 10 nsps. The ORF1b gene is located next to the ORF1a that encodes for polyprotein pp1ab which contains 16 nsps.

1.2 SARS-CoV-2 Replication Machinery

Virus entry allows the binding of angiotensin-converting enzyme 2 (ACE2) receptor and cleavage by the serine protease TMPRSS2 (in green) that allows fusion with the host membrane. Other cellular proteases, e.g., furin (in orange), facilitate pH-dependent entry through the endocytic pathway. The predominant entry routes are made to be cell type-specific and depend on the availability of selective proteases. The uncoating and release of viral RNA into the cytoplasm and translation of open reading frame 1a (ORF1a) and ORF1ab produce the two stretches of polyproteins pp1a and pp1ab respectively. These are further processed by viral proteases (encoded by ORF1a) to yield 16 nonstructural proteins. Formation of the RNA replicase–transcriptase complex (RTC) uses rough endoplasmic reticulum (ER)-derived membranes. The RTC performs the synthesis of (−) RNAs. Full-length (−) RNA copies of the genome provide templates for full-length (+) RNA genomes. Transcription further produces a subset of subgenomic RNAs, including those encoding all structural and accessory proteins. The translated structural proteins and genomic RNA are assembled into the viral nucleocapsid and envelope in the ER–Golgi intermediate compartment and are subsequently released by exocytosis.

1.3 Taxonomy of SARS-CoV-2

Since the SARS outbreak, genomic information has become ever-increasingly significant to address outbreaks caused by pathogenic coronaviruses. Before the 2019–2020 COVID-19 pandemic, there were ~1200 complete genomes of beta coronaviruses deposited in the GenBank database. The number of available genomes has increased dramatically during the pandemic, with more than 6000 complete genomes available in Genbank as of June 2019, and almost 50,000 genomic sequences in other public repositories. A variety of information including phylogenetic relationships, mode of transmission, evolutionary rates, and the role of mutations in infection and disease severity can be deduced from comparing multiple genomes. (Ugurel, Ata, and Turgut-Balik 2020)

1. 4 Analysis on Bioinformatics Tools

The program can run online on the NCBI web server. Major databases include BLASTn programs search nucleotide databases using a nucleotide query, GenBank for DNA sequences, and PubMed, a bibliographic database for biomedical literature. In our study of SARS-CoV2 for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of RNA and/or DNA sequences. It shows us a protein or nucleotide sequence (called a query) with a library or database of sequences and identifies database sequences that resemble the query sequence. (Wikipedia contributors. "BLAST (biotechnology)." Wikipedia, the Free Encyclopedia. Wikipedia, the Free Encyclopedia, 30 Dec. 2020. Web. 4 Apr. 2021). NCBI database provides FASTA sequence for human coronavirus sequence analysis which provides sequence similarity searching against protein databases. There have been identified more than 7,000 complete genome entries uploaded to Nucleotide/NCBI databases from the Coronaviridae family in the middle of 2002–2020, more than half of them are being SARS-CoV-2 and it is increasing in analogous to the expansion of sequencing technology. (Ugurel, Ata, and Turgut-Balik 2020).

2. Methods and Materials

Computational analysis of sequence alignment computer programming for bioinformatics and data management. NCBI focuses on theoretical, analytical, and applied computational approaches and widely used primary databases such as the European nucleotide archive. NCBI focuses on theoretical analytical and applied computational approaches and widely used primary databases such as the European nucleotide archive. BLAST finds regions of similarity between biological sequences. Standard Protein BLAST (BLASTP) programs search protein databases using a protein query. A FASTA sequence alignment software package is used to functional and evolutionary relationships between different sequences. Clustal is a series of widely used computer programs used in Bioinformatics for multiple sequence alignment. There have been many versions of Clustal such as CLUSTALW and CLUSTAL OMEGA, released in 1994. Open Reading Frame Finder ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter (Kumari and Choudhary 2018).

Operating system- Google chrome, MS-Windows and android

2.1 Database and Corresponding Web services

Database name	Web services type: URL
PUBMED	https://pubmed.ncbi.nlm.nih.gov/

NCBI	E-Utility web services (http://www.cbi.nlm.nih.gov
BLAST	www.ebi.ac.uk/tools/sss/ncbiblast
FASTA	www.ebi.ac.uk/tools
CLUSTAL W	https://www.genome.jp/tools-bin/clustalw
CLUSTAL OMEGA	https://www.ebi.ac.uk/Tools/msa/clustalo/
ORF FINDER	https://www.ncbi.nlm.nih.gov/orffinder/

Sequence used in Blast
Sequence 1
Nucleocapsid protein [Human coronavirus OC43]
NCBI Reference Sequence: YP_009555245.1
>YP_009555245.1 nucleocapsid protein [Human coronavirus OC43]

MSFTPGKQSSSRASSGNRSGNGILKWADQSDQFRNVQTRGRRAQPKQTATSQQPSGGNVVPYYSWFSGITQFQKGKEFEFVEGQGVPIAPGV
PATEAKGYWYRHNRRSFKTADGNQRQLLPRWYFYYLGTGPHAKDQYGTDIDGVYWVASNQADVNTPADIVDRDPSSDEAIPTRFPPGTVLPQ
GYYIEGSGRSAPNSRSTSRTSSRASSAGSRSRANSGNRTPTSGVTPDMADQIASLVLAKLGKDATKPQQVTKHTAKEVRQKILNKPRQKRSPNKQ
CTVQQCFGKRGPNQNFGGGEMLKLGTSDPQFPILAELAPTAGAFFFGSRLELAKVQNLSGNPDEPQKDVYELRYNGAIRFDSTLSGFETIMKVLNE
NLNAYQQQDGMMNMSPKPQRQRGHKNGQGENDNISVAVPKSRVQQNKSRELTAEDISLLKKMDEPYTEDTSEI

Sequence 2
Hemagglutinin-esterase [Human coronavirus OC43]
NCBI Reference Sequence: YP_009555240.1
>YP_009555240.1 hemagglutinin-esterase [Human coronavirus OC43]

MFLLPRFILVSCIIGSLGFYNPPTNVVSHVNGDWFLFGDSRSDCNHIVNINPHNYSYMDLNPVLCDSGKISSKAGNSIFRSFHFTDFYNYTGEGQQI
IFYEGVNFTPYHAFKCNRSGSNDIWMQNKGLFYTQVYKNMAVYRSLTFVNVPYVYNGSAQATALCKSGSLVLNNPAYIAPQANSGDYYYKVEA
DFYLSGCDEYIVPLCIFNGKFLSNTKYYDDSQYYFNKDTGVIYGLNSTETITTGFDLNCYYLVLPSGNYLAISNELLLTVPTKAICLNKRKDFTPVQVVD
SRWNNARQSDNMTAVACQPPYCYFRNSTTNYVGVYDINHGDAGFTSILSGLLYNSPCFSQQGVFRYDNVSSVWPLYPYGRCPTAADINIPDLPI
CVYDPLPVILLGILLGVAIVIIVVLLLYFMVDNVTRLHDA

4. Result and Discussion
Betacoronavirus 1 genome of host (Human coronavirus OC43) is single-stranded positive-sense RNA whose RefSeq protein of 496 amino acid long sequence is shown in BLAST output result to study its protein-protein interaction. The Graphical representations represent the query sequence is represented by the numbered blue bar at the top of the figure. A pairwise sequence alignment is preceded by the sequence identifier, the full definition line, and the length of the matched sequence, in amino acids. The multiple sequence alignment shows a way of arranging the sequences to identify regions of similarity that may be a consequence of evolutionary relationships between the sequences with the number of sequences submitted and the alignment score with the symbolization of different base pairs. Sequence 1 is of hemagglutinin-esterase [Human coronavirus OC43] and sequence 2 is of nucleocapsid protein [Human coronavirus OC43] where the symbol signifies an * (asterisk) indicates positions that have a single, fully conserved residue. A: (colon) indicates conservation between groups of strongly similar properties. A "." indicates a site belonging to a group exhibiting weak similarity, and the gap accounts for genetic mutations occurring from insertion or deletion in the sequence. The color code of the sequences represents the physicochemical property of the Amino acids. The analysis shows phylogeny among sequences and displays a tree known as cladogram constructed after multiple sequence alignment of all the protein sequences using Clustal omega. The ORF finder identifies all open reading frames or the possible protein-coding region in sequence. The 6 horizontal bars correspond to one of the possible reading frames. The sequence shows translation from top to bottom where regions highlighted with green are start codons, regions highlighted with red are stop codons, and regions highlighted with gray are ORFs.ORF finder output disclosing all possible open reading frames (ORFs) and their direction within the query DNA sequence details such as ORF coordinates length, strand, and frame.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per Ident	Acc. Len	Accession
ORF1a polyprotein [Human coronavirus OC43]	Human coronavi...	1032	1032	100%	0.0	100.00%	4383	YP_009924317.1

Figure 1- value closer to zero the more "significant" the match

Figure 2. *Multiple sequence alignment of Human coronavirus OC43 Genome*

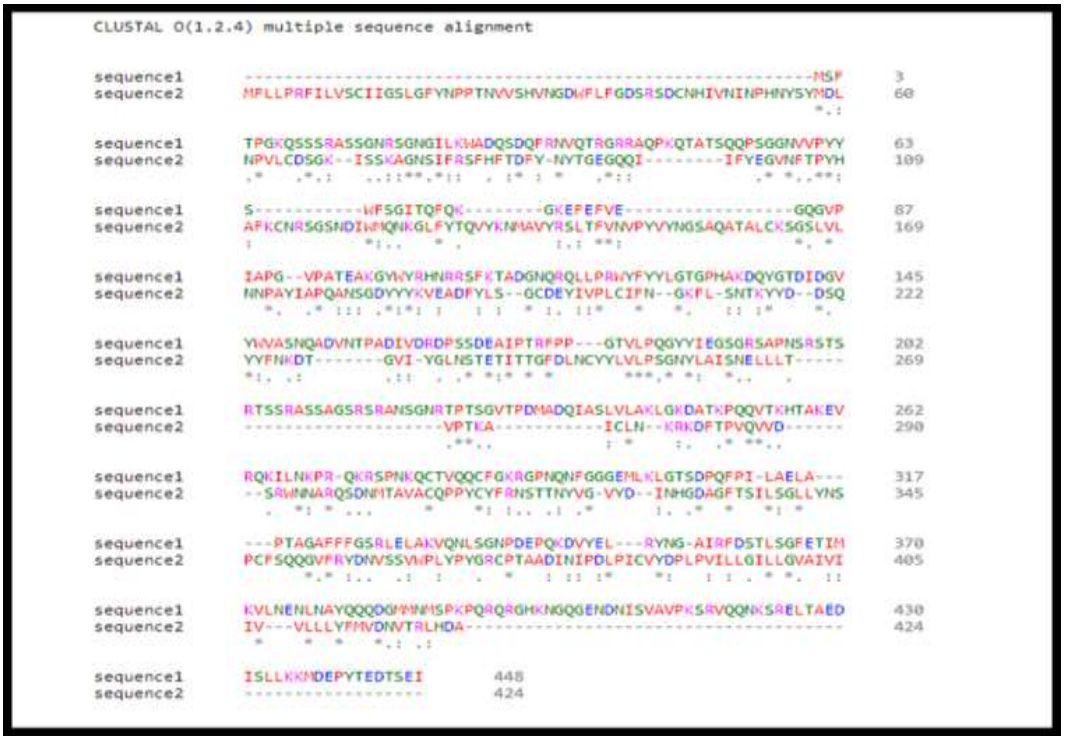

Figure 3 - *Multiple sequence alignment of Human coronavirus OCH3 Genome (color code)*

ORFs found: 12 Genetic code: 1 Start codon: 'ATG' and alternative codons

Figure 4- *ORFs in Severe acute respiratory syndrome coronavirus 2 ORF1ab*

5. Conclusion

Computational analyses to study the COVID-19 sequence in terms of protein structures, functions, phylogeny, and interactions at both molecular and sequenced levels. The BLASTP algorithm compares the protein database sequence with the query protein. E-value describes the number of hits one can "expect" to see by chance when searching a database of a particular size. The color bar in graphics summarizes the BLAST result at the top of the linear map represents a protein that matches the query sequence. Multiple sequence alignment residues provide comparative structure and functional analysis of biological sequences. The ORFs are encoded within each of the 6 translation frames 3' in the forward direction and 3' in the reverse direction so that identify the translation frame in the longest protein sequence. The open reading frame indicates the protein-coding region in an RNA sequence.

References

[1] Barati, F., Pouresmaieli, M., Ekrami, E., Asghari, S., Ziarani, F. R., & Mamoudifard, M. (2020). Potential Drugs and Remedies for the Treatment of COVID-19: a Critical Review. *Biological Procedures Online*, *22*(1), 1-17.

[2] Chen, M. J., Chang, K. J., Hsu, C. C., Lin, P. Y., & Liu, C. J. L. (2020). Precaution and prevention of coronavirus disease 2019 infection in the eye. *Journal of the Chinese Medical Association*.

[3] Forni, D., Cagliani, R., Clerici, M., & Sironi, M. (2017). Molecular evolution of human coronavirus genomes. *Trends in microbiology*, *25*(1), 35-48.

[4] Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J., ... & Yan, Y. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. *Military Medical Research*, *7*(1), 1-10.

[5] https://blast.ncbi.nlm.nih.gov/Blast.cgi

[6] https://www.cdc.gov/coronavirus/types.html

[7] https://en.wikipedia.org/w/index.php?title=COVID-19_vaccine

[8] Huang, Y., Lau, S. K., Woo, P. C., & Yuen, K. Y. (2008). CoVDB: a comprehensive database for comparative analysis of coronavirus genes and genomes. *Nucleic acids research*, *36*(Supplement_1), D504-D511.

[9] https://www.ebi.ac.uk/Tools/msa/clustalo/

[10] https://www.webmd.com/lung/covid-19-symptoms#1

[11] Kames, J., Holcomb, D. D., Kimchi, O., DiCuccio, M., Hamasaki-Katagiri, N., Wang, T., ... & Kimchi-Sarfaty, C. (2020). Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design. *Scientific reports*, *10*(1), 1-11.

[12] Kumari, U., & Choudhary, A. K. Computational Analysis of Sequences to Determine Expectation Value Commonly Used in Bioinformatics Database.

[13] La Marca, A., Capuzzo, M., Paglia, T., Roli, L., Trenti, T., & Nelson, S. M. (2020). Testing for SARS-CoV-2 (COVID-19): a systematic review and clinical guide to molecular and serological in-vitro diagnostic assays. *Reproductive biomedicine online*, *41*(3), 483-499.

[14] Llanes, A., Restrepo, C. M., Caballero, Z., Rajeev, S., Kennedy, M. A., & Lleonart, R. (2020). Betacoronavirus genomes: how genomic information has been used to deal with past outbreaks and the COVID-19 pandemic. *International journal of molecular sciences*, *21*(12), 4546.

[15] Min, Y. Q., Mo, Q., Wang, J., Deng, F., Wang, H., & Ning, Y. J. (2020). SARS-CoV-2 nsp1: bioinformatics, potential structural and functional features, and implications for drug/vaccine designs. *Frontiers in microbiology*, *11*.

[16] Raza, S., Rasheed, M. A., Zahir, W., Navid, M. T., Diwan, R. A., Awais, M., ... & Rashid, M. (2020). Structural and genetic analysis of coronaviruses spike proteins suggest pangolin as a proximate intermediate host of SARS-CoV-2 (COVID-19).

[17] Lavie, L., Medstrand, P., Schempp, W., Meese, E., & Mayer, J. (2004). Human endogenous retrovirus family HERV-K (HML-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. *Journal of virology*, *78*(16), 8788-8798.

[18] UĞUREL, O. M., ATA, O., & BALIK, D. (2020). An updated analysis of variations in SARS-CoV-2 genome. *Turkish Journal of Biology*, *44*(SI-1), 157-167.

[19] Woo, P. C., Huang, Y., Lau, S. K., & Yuen, K. Y. (2010). Coronavirus genomics and bioinformatics analysis. *viruses*, *2*(8), 1804-1820.

[20] World Health Organization. (2021). Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health, 8 January 2021.