

Analysis of Phylogenetic SARS-CoV-2 on BLAST Query using Recursive Similarity Search

Rodiah
Gunadarma University
Margonda Raya Street
Depok, Indonesia

Diana Tri Susetianingtias
Gunadarma University
Margonda Raya Street
Depok, Indonesia

Eka Patriya
Gunadarma University
Margonda Raya Street
Depok, Indonesia

ABSTRACT

One method for examining the presence of mutations in the genetic diversity of SARS-CoV involves conducting a phylogenetic tree analysis. To construct a phylogenetic tree, it is necessary to obtain suitable RNA samples. This is achieved by comparing user queries to search for sequence similarity in the NCBI database using the BLAST algorithm. It is necessary to narrow down the sample by selecting data prior to constructing a phylogenetic tree using the BLAST Query. The purpose of this action is to streamline the process of interpreting the outcomes derived from the construction of a phylogenetic tree. This study presents a novel approach for enhancing the efficiency of the BLAST Query through the utilization of four recursive searches. The optimization of BLAST queries involves selecting taxa from the experimental phylogenetic tree that exhibit sufficient dissimilarity from the input taxa. These selected taxa are then utilized as query inputs in subsequent search optimizations. Additionally, the length of the query is specifically reduced in the ORF1ab gene region. The phylogenetic tree derived from the four experiments yielded divergent interpretations, particularly in the second experiment where the taxa exhibited a high degree of homology. The optimization technique employed for BLAST queries enables the reduction of taxa samples that are to be represented in the phylogenetic tree. The present study holds the potential to serve as a valuable resource for identifying samples and constructing a phylogenetic tree for SARS-CoV-2. This, in turn, can facilitate further genetic investigations pertaining to potential viral developments, including the examination of viral mutations.

General Terms

DNA Analysis using Bioinformatics

Keywords

BLAST; ORF1a; Phylogenetic; Query; SARS-Cov-2

1. INTRODUCTION

At the end of 2019, the whole world experienced a crisis in the health sector with the spread of a new virus from the coronavirus family which was first discovered in Wuhan, China. [1] to put the world in a global scale pandemic status for quite a long time [2] namely SARS-CoV-2. The SARS-CoV-2 virus is a single RNA (Ribonucleic Acid) virus composed of structural and non-structural proteins. [3]. RNA has a major role in cell protein synthesis, converting genetic information [4] stored in DNA (Deoxyribonucleic Acid) into protein. Structurally, RNA is formed by linear polymers of four types of nucleotides, namely Adenine (A), Cytosine (C), Guanine (G), and Uracil (U), while DNA is composed of four chemical bases, namely Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). In the field of bioinformatics, these RNA or DNA sequences are commonly referred to as genome sequences [5].

Genome sequence data from various patients infected with SARS-CoV-2 are also stored in the National Center for Biotechnology Information (NCBI) database.[6]. NCBI provides datasets that are commonly used in research, especially in the medical and bioinformatics fields, from various types of living things, one of which is the SARS-CoV-2 RNA chain data in a text-based file format.[7] in GenBank or FASTA format. This genetic file format contains a sequence of nucleotide base pairs that represent the structure of a particular protein in an organism [8]. Information from the same organism may be found more than once with slight differences in the order of base pairs. These chains that have similarities can be searched for using the Basic Local Search Alignment Tool (BLAST) program. [9] web-based to compare queries from users in looking for similarity from a genetic database chain [10]. One of the methods used to analyze evolutionary changes or the possibility of mutations between these chains is the phylogenetic tree method. A phylogenetic tree is a diagram showing the branching structure of evolution [11], which is usually described by a single lineage and over time there is a branching of the lineage with a certain pattern [12], thus by forming a phylogenetic tree, the spread of a virus can be estimated, one of which is SARS-CoV-2.

Analysis of SARS-Cov-2 can be carried out by constructing a phylogenetic tree using a sample of the SARS-CoV-2 sequence that is included in the sample, to analyze variations in genetic differences between populations. Genetic distance can be calculated from the number of polymorphic base differences of a gene locus of each population based on the DNA sequence based on the similarity of characters [13] as the basis of the comparison. In biological systems, the process of evolution usually involves genetic mutations in forming new species [14]. Research [15] stated that there were 3 variants originating from the SARS-CoV-2 lineage group, namely types A, B, and C. In SARS-Cov-2, type A was identified as having the most connection with the genomes of bats and pangolins, in type B it came from a type mutation A, while type C originates from mutations in virus B. The phylogenetic method can be applied to the genome sequencing of the SARS-Cov-2 virus to help predict the future development of this virus. Phylogenetic network analysis by identifying sources of Covid-19 infection that are not recorded in spreading the infection [16][17]. The FASTA-formed SARS-CoV-2 RNA data strands can be visualized using the phylogenetic tree method which represents the relationship between taxa (processed FASTA files), where the chains that have these similarities can be searched using the program The BLAST was then aligned using the Multiple Sequence Comparison by Log Expectation (MUSCLE) program as the basis for constructing a phylogenetic tree in Biopython.

Research [17] build a phylogenetic tree using the maximum likelihood method to build a phylogenetic tree. The results of the phylogenetic formation show that the WH-Human 1 coronavirus (WHCV) is most closely related to bats, and shows 100% amino acid similarity to bat SL-CoVZC45 on nsp7 and E protein, thus indicating that bats are possible hosts for the reservoir. WHCV virus. Research [18] visualize phylogenetic trees from genome sequences obtained from two tourists from China who visited Rome and one person from Italy. Research [19] perform a search for similar sequences using BLAST and several variants of the BLAST program. Research [20] forming a phylogenetic tree by further exploring the ORF1ab gene in the SARS-CoV-2 virus and its proximity to other virus variants, as well as in research [21] which states that the nucleotide strand of the ORF1ab gene in the SARS-CoV-2 virus is quite long.

On research [17] [18] related to the analysis of the occurrence of viral mutations by looking at the formation of phylogenetic trees has not explained the processes that occurred before the phylogenetic trees were formed, especially in the collection of RNA data that has similarity with the research object using BLAST. The database at NCBI is large enough to store various genetic information from various types of organisms and is inseparable only for the SARS-CoV-2 virus. Refinement of samples using BLAST Query is very necessary in forming phylogenetic trees, so that the results of forming phylogenetic trees can provide a more in-depth interpretation. As with other search engines, to get something relevant, you need keywords that adequately represent the search results. This study proposes a Query optimization process in a specific BLAST program using the ORF1ab gene as a query parameter in four search trials. The results of each of these experiments were then visualized using a phylogenetic tree using the MUSCLE and BioPython programs. The SARS-CoV-2 viral RNA strand in research [17] previously used as the first trial query parameter, and then recursively using one of the other RNA strands formed in the phylogenetic tree as input for the next query experiment. Several samples formed in each experiment were put back together into a single phylogenetic tree in the hope of providing a better interpretation compared to the first experiment as well as further research as a reference and deeper analysis in interpreting the SARS-CoV-2 virus phylogenetic tree.

2. RESEARCH METHOD

The method in this study consists of several process stages to optimize the BLAST Query parameters. The research was carried out by first extracting data on the SARS-CoV-2 RNA strand which is formed by the arrangement of nucleotides and in each particular sequence represents the Gene ID. Similar RNA strands can then be searched from the NCBI database using BLAST so that a phylogenetic tree can be formed for analysis using the Multiple Sequence Alignment and UPGMA methods. Fig. 1. shows the general chart of the methods in this study.

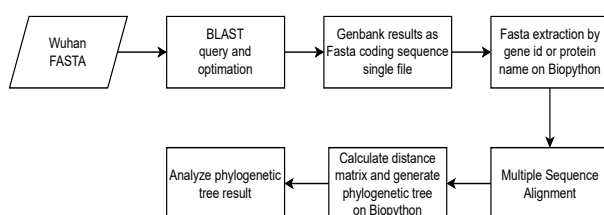


Fig 1 : General Chart of Research Methodology

As can be seen from the Fig. 1, this research use SARS-CoV-2 RNA NC_045512.2 with FASTA format[6]. A phylogenetic tree will be formed using the help of MUSCLE and BioPython describe which sequence has the most distant branch length to be reused as search query optimization in subsequent experiments. Target of this research is to look for sequences that are similar to the sequence with id NC_045122.2 in the ORF1ab gene.

2.1 Wuhan FASTA Dataset

The data used in this study is SARS-CoV-2 RNA with the accession number NC_045512.2 which is published publicly on the NCBI website [6]. The RNA strand is 29903 bp long and was withdrawn in 2019 from a 41-year-old male patient who had no history of diabetes, tuberculosis or hepatitis [17]. The RNA information can then be retrieved in a text file with FASTA format either in the form of a whole strand or broken down based on the Gene ID (Coding Sequence) as can be seen in Fig. 2.

```

>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wu
genome
ATTAAAGGTTTATACCTTCCAGGTAACAACCAACCACTTCGATCTCTGTAGATCTGTTCTTAA
CGAAGCTTTAAATCTGTGTGGCTGTCACTCGGTGATCTTAGTCACTACGACAGTATAATTAAT
TAATTACTGTCTGTGACAGGACACGAGTAAGTCTGTCTATCTTGTGACGGCTGCTACGGTTT
TTGAGCCGATCATCAGCACATCTAGGTTTCTGCGGGTGTGACCGAAAGGTAAAGTGGAGACCT
CCTGGTTTCAACGAGAAAAACACGTCCTCACTAGTTTGCCTGTTTACAGGTTTGGAGCTGCT
GTGGCTTTGGAGACTCCGTGGAGGAGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCA
CTTAGTAGAAGTTGAAAAAGGCGTTTTCCTCACTTGAACAGCCCTATGTTCATCAAGCTTCG
GATGCTCGAAGTGCACCTCATGGTATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCAAT
GTAGTGGTGAGACACTTGGTGCTTGTCCCTCATGTGGGGCAAAATACAGTGGCTTACCGCA
AGGTTCTCTTCGTAAGAAGCGTAATAAAGGAGCTGGTGGCCATAGTTACGGCCGATCAAGG
GCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAACTGGAACACTAATACATAG
CGTGGT
  
```

Fig.2 : Fasta File NC_045512.2

2.2 Optimization on BLAST Query

The main input in this study was WHCV SARS-CoV-2 [17] which has one of the polypeptides with Gene ID ORF1ab which is also shared by other coronaviruses [22]. The ORF1ab gene ID in the WHCV SARS-CoV-2 strand itself is quite long compared to other polypeptides so that it can be used as a reference to look for fairly homologous similarity. [23]. The location of ORF1ab base pairs can be used as a subquery parameter on the BLAST page to narrow down the number of words used in the mega blast program. BLAST will look for the genome strand with the highest similarity to be displayed in the inquiry process [24]. For each BLAST query search result, a phylogenetic tree will be formed using the help of MUSCLE and BioPython as shown in Fig. 3.

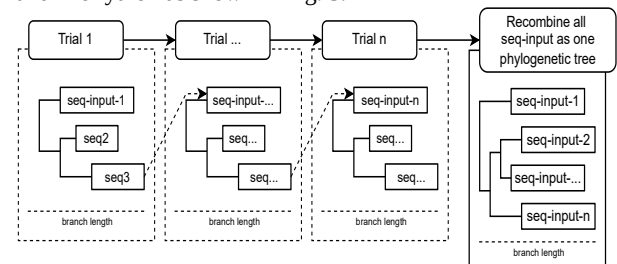


Fig. 3 : Searching on BLAST is Optimized Sequentially With The Help Of MUSCLE To Build Phylogenetic Trees

2.3 Withdrawing Query Results into Fasta Format

The BLAST search results can be pulled into a single file in FASTA format containing the complete sequence containing the complete nucleotides of the search results as shown in Fig. 4. In good cases, each record can also be downloaded into a coding sequence format which has broken one sequence into several sequences based on the code defined by the researcher who uploaded the sequence. The target of this research is to look for sequences that are similar to the sequence with id NC_045122.2 in the ORF1ab gene.

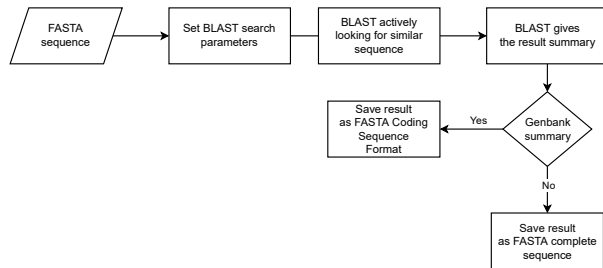


Fig. 4 : Search results on BLAST in Coding Sequence

2.4 Subset File Fasta

BioPython has the capability to process FASTA files programmatically, in this study BioPython is used to extract FASTA files based on the protein code or prune them based on specific index range using the SeqIO Function using the following pseudocode snippet:

```

def extract_protein_name(seq):
    description = seq.description
    x = re.findall("[protein=([a-zA-Z\s0-9]+)]", description)
    return x[0]

def filter_records_by_protein(arr, protein_name):
    arr_new = list()
    for records in arr:
        records = list(filter(lambda x:
            (extract_protein_name(x) == protein_name), records))
        for record in records:
            arr_new.append(record)
    return arr_new
  
```

The SeqIO function in Fig. 5. will filter sequences that contain certain proteins in the FASTA definition according to the parameters of the protein name that is input



Fig 5 : FASTA Coding Reading and Filtering Algorithm Sequence with ORF1ab Protein

The SeqIO function will then perform pruning by narrowing the length of the FASTA file being analyzed based on the location of the nucleotide range as can be seen in Fig. 6.

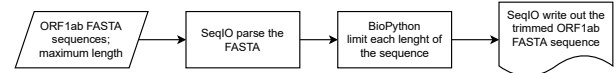


Fig. 6 : FASTA Coding Reading and Filtering Algorithm Sequence with ORF1ab Protein

This pruning will produce information as can be seen in Fig. 7(A) The information indicating the sequence constitutes the ORF1ab protein and Fig. 7(B) If in FASTA it is cut specifically in the boxed area then this one sequence will only have a length of 23 nucleotides.

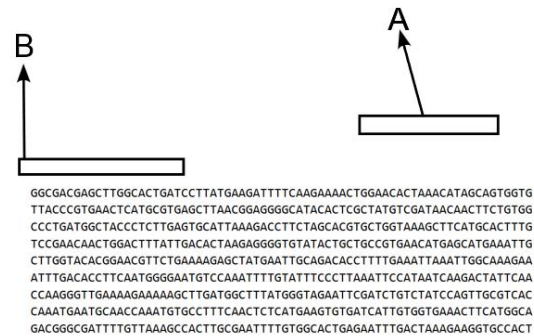


Fig. 7 : FASTA with Coding Sequence File

2.5 Multiple Sequence Alignment (MSA) with MUSCLE

FASTA files that contain genome strands that were previously pulled from BLAST query search results can be aligned by aligning the FASTA files which in the MUSCLE program use the Parallel Perturbed Probcons algorithm as the main algorithm, namely re-implementation of Probcons with several improvisations including parallelization, parameter perturbation, and support for strands of amino acids and nucleotides [25]. FASTA files that have been aligned from MSA will be used as input to the BioPython module using the AlignIO module. The phylogenetic tree of MSA objects that has been formed before can be constructed by calculating the Distance Matrix. The Distance Matrix used in phylogenetics depends on the size of the genetic distance between the sequences being classified, so it takes MSA as input. The use of the Distance Matrix in building an all-to-all matrix from a series of sequences Queries that describe the distance between each pair of sequences are carried out in stages [26] :

1. Calculating pairwise distances: Genetic distances between all pairs of taxa based on the order of concordance or trait comparison
2. Build a distance matrix: Build a matrix that represents the pairwise distances between taxa.
3. Construct a tree using Neighbor-Joining to build a tree based on a distance matrix which will iteratively combine the closest pair of taxa to form a tree, with branch lengths proportional to the distance.

Each FASTA strand that is used iteratively in the BLAST search is reassembled separately into one FASTA file and the process of forming the same phylogenetic tree is carried out based on the previous steps with the expectation that the furthest evolution of the WHCV sequence strand will be formed [17] as can be seen in Fig. 8.

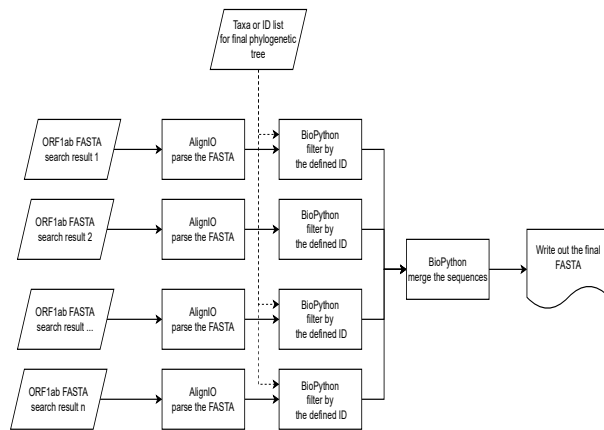


Fig. 8 : Merging of taxa into one FASTA

3. RESULT AND DISCUSSION

The results of the study were carried out using the Wuhan SARS-CoV-2 FASTA file [17] with SeqID NC_045512.2 by optimizing the query parameters in the BLAST program in forming a sequential phylogenetic tree.

3.1 Test Results Experiment 1: BLAST query using the whole sequence NC_045512.2

The first test was carried out on the genome sequence with SeqID NC_045512.2 the BLAST program, as can be seen in Table 1.

Table 1. Results of Trial 1 Query Optimization

Searching Parameter	Value
Input	NC_045512.2
SubQuery Interval	null (whole sequence)
Hitlist Size	100 record

An error occurred in the search which resulted in CPU usage exceeding the limit and was calculated manually with the execution time in this experiment taking around 1 minute 44 seconds. As can be seen in Fig. 9.

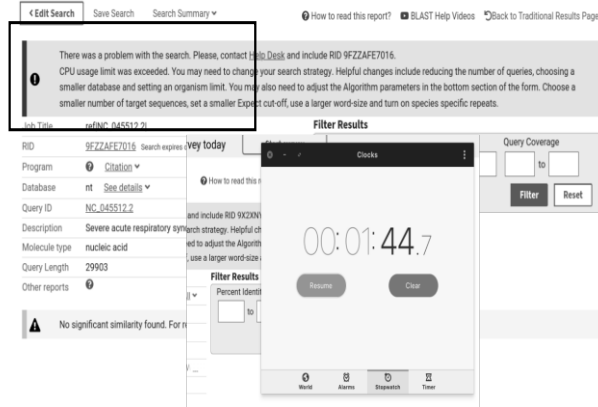


Fig 9 : Trial 1 CPU Usage Notification

As can be seen from the Fig. 9, The system also provides suggestions for re-optimizing parameters by setting up the Query used in the BLAST search.

3.2 Test Results for Experiment 2: BLAST query using the whole sequence NC_045512.2

Failure in the data search process using the whole genome parameter NC_045512.2 which was run in the first experiment, then a second experiment was carried out on the Query parameter which was optimized again by cutting the Query length, as can be seen in Table 2. Query optimization is done by defining the starting and ending locations of the queries used from the whole sequence. One ID gene that has a relatively large sequence length is ORF1ab [27] which would be quite homologous if referenced to this gene. The sequence NC_045512.2 has the ORF1ab gene located at positions 266-21555 [28] which will be used as the initial and final Query indexes in the BLAST program with a limit of 50 items [29].

Table 2. Results of Trial 2 Query Optimization

Searching Parameter	Value
Input	NC_045512.2
SubQuery Interval	266..21555
Hitlist Size	50 record

The search results in the second experiment were then downloaded into the coding sequence format, out of a total of 50 records, only 38 records had coding sequence information combined with the NC_045512.2 input file, a total of 39 records which will be processed further. The coding sequence information is quite helpful in knowing the location of a particular gene ID on an RNA strand. In building a phylogenetic tree, it is necessary to first align the 39 records using the MUSCLE program, as shown in Fig. 10. The MUSCLE program failed to synchronize the record and was forced to stop by the CPU after being executed for 12.56 seconds.

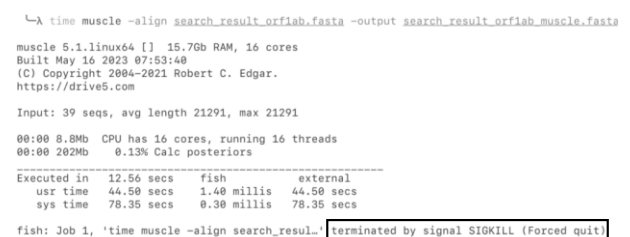


Fig 10: MUSCLE Can Not Perform Task On Very Large Records

Each sequence in the ORF1ab gene ID can be over 20000 nucleotides long [6] multiplied by a total of 39 records. In this experiment, another way for MUSCLE to carry out its task of aligning a total of 39 records containing ORF1ab information is to cut the nucleotide length of each record as can be seen in Fig. 11.

```
[#] x time muscle -align search_result_orf1ab_1000.fasta -output search_result_orf1ab_1000_muscle.fasta

muscle 5.1.linux64 [ ] 15.7Gb RAM, 16 cores
Built May 16 2023 07:53:40
(C) Copyright 2004-2021 Robert C. Edgar.
https://drive5.com

Input: 39 seqs, avg length 1000, max 1000

00:00 7.5Mb CPU has 16 cores, running 16 threads
00:09 1.1Gb 100.0% Calc posteriors
00:09 1.1Gb 100.0% Consistency (1/2)
00:10 1.1Gb 100.0% Consistency (2/2)
00:10 1.1Gb 100.0% UPGMA5
00:10 1.2Gb 100.0% Refining

-----
Executed in 10.27 secs fish external
usr time 126.50 secs 315.00 micros 126.50 secs
sys time 8.83 secs 248.00 micros 8.83 secs
```

Fig 11 : MUSCLE Can run Tasks with Records that have been trimmed

This research using BioPython's help, every 39 records are cut, only taking the first 1000 nucleotides, counting from the first location of the ORF1ab position, each record is located. Thus MUSCLE is able to align these sequences with a terminal execution time of 10.27 seconds.

3.3 Test Results for Experiment 3: BLAST query using the subquery sequence OM065360.1

The results of the two experiments have not shown a more meaningful phylogenetic tree because each node formed shows a branch length of 0 [30], where the longer a branch length (viewed from left to right) indicates the greater the changes that occur to form the knot [31]. Table 3 is a query optimization in experiment 3.

Table 3. Result of Trial 3 Query Optimization

Searching Parameter	Value
Input	OM065364.1
SubQuery Interval	1000-2000
Hitlist Size	50 record

However, in the visually constructed phylogenetic tree, the taxa that is most branched towards SeqID NC_045512.2 is SeqID OM065364.1 [32] which will be used as a query in this experiment. Queries that are used as parameters are cut at indexes 1000-200 to see any differences in the query results and the resulting phylogenetic tree. Like the previous process, the BLAST query results in the third experiment were also downloaded in FASTA coding sequence format with a total of 50 + 1 records (1 record is the input query) and extracted using BioPython. In coding the ORF1ab protein sequence, the phylogenetic tree will be formed by first aligning it with the MUSCLE program. MUSCLE cannot process files with too long sequence lengths, thus in subsequent experiments this process will be skipped and immediately cut specifically for the first 1000 nucleotide lengths calculated from the first location of the ORF1ab sequence coding.

Based on the search results in the third experiment, a phylogenetic tree was formed in 51 records of the third experiment, which consisted of two large clusters namely the USA/CAV/VA and USA/NM-UNM-TRL clusters as can be seen in Fig. 12.

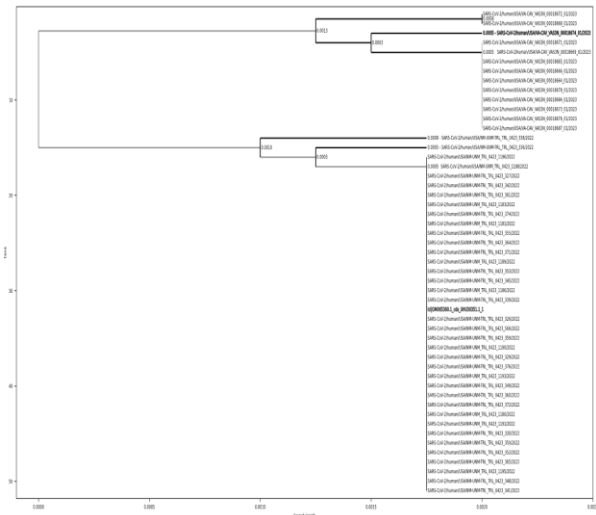


Fig 12 : Experiment 3 Phylogenetic Tree

In the phylogenetic tree formed in Fig. 12. Form a large cluster with the code USA/NM-UNM-TRL (New Mexico) and USA/CAV/VA (Virginia) where the query that is used as input with ID OM065360.1 is in the New Mexico cluster. In the second largest cluster branch which is colored light blue, there is one taxa branch relative to the branch [32] with SeqID OR184938.1 and used as a query parameter in the next experiment. In constructing the final phylogenetic tree, two taxa relative to the taxa that were queried in the next experiment, as well as two taxa that were relative through two vertices of the taxa that were used as the query in this experiment were used.

3.4. Test Results of Experiment 4: BLAST Query using the subquery sequence OR184938.1

In the fourth experiment, one of the sequences with similarity far enough from the input of the third experiment was used as a BLAST query parameter, namely SARS-CoV2/human/USA/VACAV_VAS3N_00018674_01/2023 with SeqID OR184938.1 with query parameters as can be seen in Table 4.

Table 4. Result of Trial 4 Query Optimization

Searching Parameter	Value
Input	OR184938.1
SubQuery Interval	2000-3000 (located on ORF1ab area)
Hitlist Size	50 record

With the failure of two attempts at alignment of ORF1ab sequences in the second and third trials, specifically in this experiment the sequences that had been extracted specifically in the ORF1ab gene were immediately trimmed in the first 1000 nucleotide sequences to be aligned in the MUSCLE program to produce phylogenetic trees for the next 51 experimental records, visualized record as can be seen in Fig. 13.

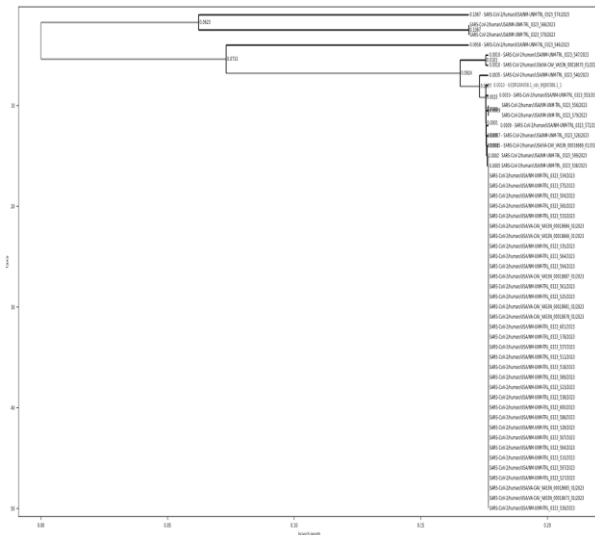


Fig 13 : Experiment 4 Phylogenetic Tree

The phylogenetic tree for the next 51 experimental records, visualized that record OR184938.1 (USA, Virginia) is topologically very close relative to the 42 records, of which 36 records have branch length 0 forming a cluster, and 6 records in an- all of them have a branch length that is very close to 0. The other closest remaining taxa are followed by the following taxa which are then used as taxa in the formation of the next phylogenetic tree. Using the BioPython program, the three experiments were re-selected into one FASTA file to construct the phylogenetic tree.

Fig. 14 is a phylogenetic tree formed from the results of re-selection into one FASTA file from the second experiment, third experiment, fourth experiment and on the Thick line are the taxa used as query parameters in that experiment. It can be seen that the first sequence used as input NC_045512.2 is located in the same cluster as the two sequences selected in the second experiment and several sequences in the third experiment, where the sequence used as input in the third experiment OM065360.1 (which previously also appeared in second trial search) are located in the same cluster, even though in the phylogenetic tree formed in Fig. 11. they are located very far away through several clusters. The sequence with ID OM065360.1 is also located in a cluster that is very close to the sequence ID OR184938.1 which was used as input in the fourth experiment, where the sequence ID OR184938.1 is located in a different cluster in the phylogenetic tree.

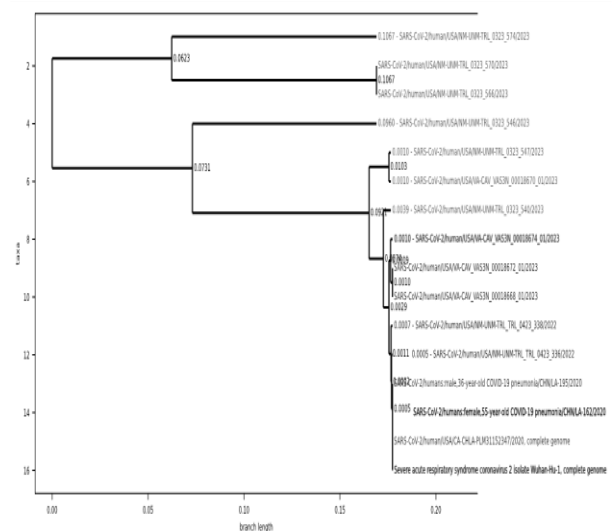


Fig. 14. Phylogenetic Tree of Selection Results
Back to Fasta Files

From this last process, it can be seen that optimizing the BLAST query using taxa that have appeared in previous experiments produces phylogenetic trees with quite different interpretations.

4. CONCLUSION

Based on the test results by conducting experiments to optimize the BLAST query in forming a phylogenetic tree, including: The first experiment failed because the query used in the search was too long so it needed to be trimmed by narrowing the query to the ORF1ab gene area, namely 266-21555. The third and fourth trials involved one of the taxa formed in the previous experiment and reduced the length of the query by about 1001. The phylogenetic tree was built with the help of the MUSCLE program and in the second experiment the sequence alignment task was not carried out successfully due to the large amount of data being processed so it needed to be trimmed first, in this study it is taken the first 1000 nucleotides counted from the first location of ORF1ab. Optimization in this experiment was carried out by using RNA chains which seemed to have quite a similarity to the RNA chains used as input in the BLAST search shown in the visualization of the phylogenetic tree and narrowing down the length of the query. In each search iteration, the combined results of several RNA chains from each experiment resulted in quite different visualizations that opened up space for analysis compared to the phylogenetic tree built in the second experiment.

Further research can be carried out by analyzing more deeply the parameters used in the BLAST program such as gap parameters and scoring and their impact on search results, making deeper observations in analyzing the structure between the RNA strands of the SARS-CoV-2 virus so that other information is obtained such as the possibility of mutations. on certain proteins and further observations on other genes or nucleotide locations belonging to the RNA chain of the SARS-CoV-2 virus.

5. ACKNOWLEDGMENTS

Thank you Directorate of Research, Technology and Community Service, Directorate General of Higher Education, Research and Technology Ministry of Education, Culture, Research and Technology. Thank you for the support of the Gunadarma University Research Institute for the smooth implementation of this research.

6. REFERENCES

- [1] A. A. Balkhair, "Covid-19 pandemic: A new chapter in the history of infectious diseases," *Oman Med. J.*, vol. 35, no. 2, pp. 2–3, 2020, doi: 10.5001/OMJ.2020.41.
- [2] J. Hiscott *et al.*, "The global impact of the coronavirus pandemic," *Cytokine Growth Factor Rev.*, vol. 53, no. May, pp. 1–9, 2020, doi: 10.1016/j.cytogfr.2020.05.010.
- [3] S. B. Kadam, G. S. Sukhramani, P. Bishnoi, A. A. Pable, and V. T. Barvkar, "SARS-CoV-2, the pandemic coronavirus: Molecular and structural insights," *J. Basic Microbiol.*, vol. 61, no. 3, pp. 180–202, 2021, doi: 10.1002/jobm.202000537.
- [4] S. Minchin and J. Lodge, "Understanding biochemistry: Structure and function of nucleic acids," *Essays Biochem.*, vol. 63, no. 4, pp. 433–456, 2019, doi: 10.1042/EBC20180038.
- [5] D. Skanderson, M. Meauley, and J. O. E. Garrett, "M a n a g i n g t h e Risk of —," vol. 53, no. March, pp. 74–80, 2015.
- [6] N. C. for B. I. (NCBI), "No Title," *NCBI SARS-CoV-2 Resources*, 2023. <https://www.ncbi.nlm.nih.gov/sars-cov-2/>
- [7] P. V. Markov *et al.*, "The evolution of SARS-CoV-2," *Nat. Rev. Microbiol.*, vol. 21, no. June, 2023, doi: 10.1038/s41579-023-00878-2.
- [8] S. Chatterjee *et al.*, "Whole-genome analysis and mutation pattern of SARS-CoV-2 during first and second wave outbreak in Gwangju, Republic of Korea," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, 2022, doi: 10.1038/s41598-022-14989-y.
- [9] K. J. Menlove, M. Clement, and K. A. Crandall, "Chapter 1 Similarity Searching Using BLAST," no. 1, pp. 1–22, doi: 10.1007/978-1-59745-251-9.
- [10] K. C. Samal, J. P. Sahoo, L. Behera, and T. Dash, "Understanding the BLAST (Basic Local Alignment Search Tool) Program and a Step-by-step Guide for its use in Life Science Research," *Bhartiya Krishi Anusandhan Patrika*, vol. 36, no. Of, pp. 55–61, 2021, doi: 10.18805/bkap283.
- [11] A. E. Gorbalenya, "Phylogeny of Viruses ☆," no. December 2016, pp. 1–7, 2017.
- [12] D. Dylus, A. Altenhoff, S. Majidian, F. J. Sedlazeck, and C. Dessimoz, "Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree," *Nat. Biotechnol.*, 2023, doi: 10.1038/s41587-023-01753-4.
- [13] D. J. Witherspoon *et al.*, "Genetic similarities within and between human populations," *Genetics*, vol. 176, no. 1, pp. 351–359, 2007, doi: 10.1534/genetics.106.067355.
- [14] D. Singh and S. V. Yi, "On the origin and evolution of SARS-CoV-2," *Exp. Mol. Med.*, vol. 53, no. 4, pp. 537–547, 2021, doi: 10.1038/s12276-021-00604-z.
- [15] I. Syafarina, T. Wirahman, S. B. Iryanto, and A. L. Latifah, "Epidemic Data Analysis of Three Variants of COVID-19 Spread in Indonesia," *J. Ilmu Komput. dan Inf.*, vol. 15, no. 1, pp. 47–54, 2022, doi: 10.21609/jiki.v15i1.1055.
- [16] P. Zhou *et al.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, pp. 270–273, 2020, doi: 10.1038/s41586-020-2012-7.
- [17] F. Wu *et al.*, "A new coronavirus associated with human respiratory disease in China," *Nature*, vol. 579, no. 7798, pp. 265–269, 2020, doi: 10.1038/s41586-020-2008-3.
- [18] P. Stefanelli *et al.*, "Whole genome and phylogenetic analysis of two SARSCoV-2 strains isolated in Italy in January and February 2020: Additional clues on multiple introductions and further circulation in Europe," *Eurosurveillance*, vol. 25, no. 13, pp. 1–5, 2020, doi: 10.2807/1560-7917.ES.2020.25.13.2000305.
- [19] G. Hu and L. Kurgan, "Sequence Similarity Searching," *Curr. Protoc. Protein Sci.*, vol. 95, no. 1, pp. 1–19, 2019, doi: 10.1002/cpps.71.
- [20] G. Li, L. Zhang, and N. Du, "Relative synonymous codon usage of ORF1ab in SARS-CoV-2 and SARS-CoV," *Genes and Genomics*, vol. 43, no. 11, pp. 1351–1359, 2021, doi: 10.1007/s13258-021-01136-6.
- [21] M. Fahmi, Y. Kubota, and M. Ito, "Nonstructural proteins NS7b and NS8 are likely to be phylogenetically associated with evolution of 2019-nCoV," *Infect. Genet. Evol.*, vol. 81, no. March, p. 104272, 2020, doi: 10.1016/j.meegid.2020.104272.
- [22] S. Srinivasan *et al.*, "Structural Genomics of SARS-CoV-2 Indicates," *Viruses*, vol. 12, no. 360, pp. 1–17, 2020.
- [23] E. Sheikhzadeh, S. Eissa, A. Ismail, and M. Zourob, "Diagnostic techniques for COVID-19 and new developments," *Talanta*, vol. 220, no. May, 2020, doi: 10.1016/j.talanta.2020.121392.
- [24] G. F. Webster, T. Poyner, and B. Cunliffe, "Clinical review," *Bmj*, vol. 325, no. 7362, p. 475, 2002, doi: 10.1136/bmj.325.7362.475.
- [25] R. C. Edgar, "Sequence alignment replicates reveal errors in molecular phylogenies," pp. 1–17, 2021.
- [26] B. P. Blackburne and S. Whelan, "Measuring the distance between multiple sequence alignments," vol. 28, no. 4, pp. 495–502, 2012, doi: 10.1093/bioinformatics/btr701.E. C. R. Id, J. H. Chariker, and D. Chung, "Variant analysis of 1, 040 SARS-CoV-2 genomes," no. Fig 2, pp. 1–18, 2020, doi: 10.1371/journal.pone.0241535.
- [27] T. T. Nguyen, P. N. Pathirana, T. Nguyen, H. Nguyen, A. Bhatti, and C. Dinh, "Genomic Mutations and Changes in Protein Secondary Structure and Solvent Accessibility of SARS-CoV-2 (COVID-19 Virus)," vol. 2, 2020.
- [28] N. Sharif, "Phylogenetic and whole genome analysis of first seven SARS-CoV-2 isolates in Bangladesh Graphical abstract :," vol. 63, no. 2003, 2020.
- [29] M. Binet, O. Gascuel, C. Scornavacca, E. J. P. Douzery, and F. Pardi, "Fast and accurate branch lengths estimation for phylogenomic trees," *BMC Bioinformatics*, pp. 1–19, 2016, doi: 10.1186/s12859-015-0821-8.
- [30] B. A. Yakici, H. A. Ogilvie, and L. Nakhleh, "Phylogenetic Network Dissimilarity Measures That Take Branch Lengths Into Account," 2022.
- [31] N. C. for B. I. (NCBI), "No Title," *Gen Bank*, 2020, [Online]. Available: <https://www.ncbi.nlm.nih.gov/nuccore/OM065360>