Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands Technical Aspects Report

Franz Dizon,, Sidney Guaro, Syd De Ama
College of Computer Studies
De La Salle University
{franz_dizon, sidney_guaro, syd_deama}@dlsu.edu.ph

Abstract—**In this report we examine the tools used and what are the underlying theories behind it. We also provide a pipeline for the extension of the paper.**

Keywords—Bioinformatics, Phylogenetic analysis, BEAST analysis

I. Introduction

During the COVID-19 epidemic, preventing the transportation of infectious people has become an essential intervention. It is important to understand the source of infection in order to initiate effective and successful contact tracing for SARS-CoV-2 infected patients and quarantined persons exposed to infectious patients. For COVID-19 patients whose source of infection cannot be fully understood, it will be resolved by phylogenetic analysis of the isolated virus.

Phylogenetic analysis is often considered a daunting and complex process. In this report, we will discuss how [1] constructs a phylogenetic tree to have wise public health decisions.

II. Objectives

The report objectives will be:

- To understand how the phylogenetic tree is formed and analyzed.

For the first part of our presentation, we will focus on understanding the technical aspects behind the phylogenetic tree in [1]. This is to have a comprehension on how the sequences are used to construct a phylogenetic tree in order to have an understanding of how the viruses evolve.

- To discuss the plan on extending the paper.

For the second part of our presentation, we will discuss the plans on how we can extend the [1] paper. This will be presented in a pipeline..

III. Sequence Data Analysis

A. Demultiplex

When NGS is performed on a set of samples, unlike conventional methods, the samples are combined all into one and load it into the sequencer. The mixing up of samples (libraries) is "Multiplexing". When libraries for NGS are prepared, two oligonucleotides to the DNA fragments are added - Adapters and index. Adapters are oligonucleotides that enable DNA fragments to attach to the flowcell. Indexes or barcodes help to differentiate samples after sequencing. For example, imagine having 10 different samples, during library preparation 10 different indexes will be added to the 10 samples. Before loading on to the sequencer, all the libraries will be combined together. During sequencing, the index sequences are also sequenced. After sequencing, the fastq file will contain all the sequences from the 10 libraries. Now, the sequences have to be "De-multiplex" from the single fastq file to obtain the sequences from 10 different samples. With a computational tool, sequences will be sorted with the same index to a group and according to the index used. In [1], demultiplexed sequences are aligned with the reference sequence using minimap2 [4].

B. Homopolymer

Different sequencing platforms produce several types of errors in sequencing, which may cause variants named incorrectly. The determination of nucleotides, homopolymeric regions, is the most frequent cause of sequencing errors across platforms. These are areas that have the same nucleotide stretches (e.g. AAAAA or TTTTTTTT). This allows variant-callers to recognize insertions and deletions not explicitly present in the sample inside homopolymeric regions. In [1], the homopolymeric region in the genomes with a coverage of <30 were replaced by a 'N'.

IV. Phylogenetic analysis

To understand the evolutionary relationship of the SARS-CoV-2 in the Netherlands and in the world where sequences are uploaded in GISAID, the paper used MUSCLE (MUltiple Sequence Comparison by Log-Expectation) [2] to align the sequences in Netherlands and in the GISAID. Afterwards, [1] used IQ-Tree [7] to construct the Phylogenetic tree and BEAST [8] to construct the Bayesian phylogenetic tree.

A. Phylogenetic Tree Overview

Before we discuss how to generate the phylogenetic tree, it would be more reasonable to discuss the structure of the phylogenetic tree to expect.

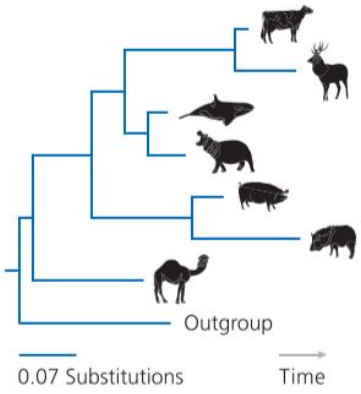

Fig. 1: A sample phylogenetic tree of artiodactyls (whale being on it is still under discussion).
Source: Adapted from [5]

A phylogenetic tree, sometimes known as evolutionary tree, is a tree diagram showing the history of divergence and evolutionary change of organisms, and it depicts its relationship with each other [5]. Each node in a tree represents a species. The root node is the common ancestor of all species represented by terminal and non-terminal nodes, and a terminal node is the common ancestor of all the species in its subtree (also known as clade). The splitting of branches, or the internal edges of the tree, indicates divergence of one kind of species, to two or more kinds of species. Note that if there are more than two branches coming from a parent node, it indicates uncertainty, it means that there is a lack of evidence to conclude which species comes first. A sample of a phylogenetic tree is shown on Fig. 1. In our case of phylogenetic tree inferencing, the branch lengths on the x-axis represent the nucleotide substitutions per site. Note that not all methods of phylogenetic tree inferencing have this feature, some of the more basic inferencing methods give no importance on branch lengths.

B. Multiple Sequence Alignment using MUSCLE

In [1], the SARS-CoV-2 genomes were retrieved from GISAID on March 22, 2020, and multiple sequences were aligned with the Dutch SARS-CoV-2 sequences using MUSCLE. They excluded sequences with >10% 'N's and manually checked the aligned-sequences for discrepancies.

MUSCLE aligns the sequences iteratively to achieve high accuracy as compared to progressive alignments used in other multiple alignment sequence software like CLUSTAW [2]. Fig. 1. Shows the flow of MUSCLE.

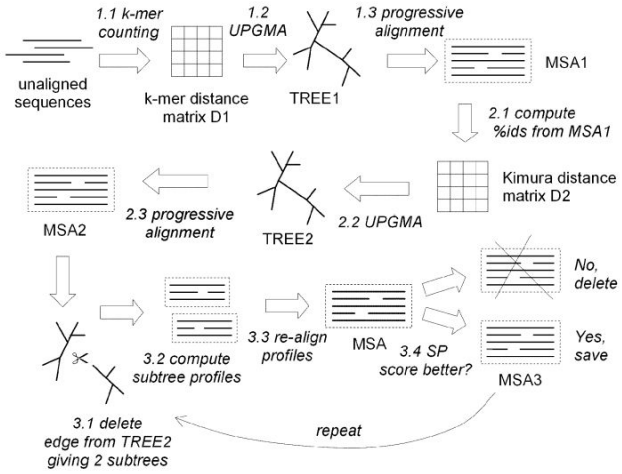

Fig. 2: Diagram of MUSCLE multiple sequence alignment. There are three main stages: Stage 1 (draft progressive), Stage 2 (improved progressive) and Stage 3 (refinement). At the completion of each stage, a multiple alignment is available in which the algorithm may terminate.
Source: Adapted from [3]

MUSCLE flow discussion as in [3]:

Stage 1, Draft progressive. In this stage, the goal is to produce multiple alignment, emphasizing speed over accuracy. This stage builds a progressive alignment.

1.1 The kmer distance is computed for each pair of input sequences, giving distance matrix D1.

In this step, using k-mer counting, the similarity of each pair of sequences is computed indicating their degree of evolutionary divergence [7]. K-mer counting counts the number of subsequences occurring in the sequence. In [1], they used the default k-mer option, 6-mer. Note that k-mers with no matches is disregarded.

To construct the distance matrix D1, MUSCLE define the following similarity measure between sequences X and Y [3]:

$$F = \Sigma_{\tau} \min [n_X(\tau), n_Y(\tau)] / [\min (L_X, L_Y) - k + 1]. \quad (1)$$

Here τ is a k-mer, L_X, L_Y are the sequence lengths, and $n_X(\tau)$ and $n_Y(\tau)$ are the number of times τ occurs in X and Y respectively. This definition can be motivated by considering an alignment of X to Y and defining the similarity to be the fraction of k-mers that are conserved between the two sequences. The denominator of F is the maximum number of k-mers that could be aligned. The definition of F is an approximation in which it is assumed that common k-mers are always alignable to one another (after correcting for excesses) [2]. Below shows an example of similarity measure using k-mer counting.

Sequence A: ATGGTGTGAGAACTTTTGTATAACTTC

Sequence B: GTGGTGTAGGAACTTTGGTCTGGTGT_
Sequence C: TTGGTGTAATAACTTTTTGCGT_____
Sequence D: CTGGTGTGCAAACTTTGAGT_____
Sequence E: GTGGTGTGCCAACTTTACGAAACTTC_
Sequence F: GTGGTGTTAGAACTTTTAACCGGCCGT

TABLE 1.1
Example output of k-mer Counting

Sequence name	Sequences		
	AACTTC	TGGTGT	AACTTT
Seq. A	1	1	1
Seq. B	0	2	1
Seq. C	0	1	1
Seq. D	0	1	1
Seq. E	1	1	1
Seq. F	0	1	1

TABLE 1.2
Example output of similarity measure using Equation 1

Sequence name	Sequences					
	Seq. A	Seq. B	Seq. C	Seq. D	Seq. E	Seq. F
Seq. A	-		-	-	-	-
Seq. B	0.09524	-	-	-	-	-
Seq. C	0.11765	0.11765	-	-	-	-
Seq. D	0.13333	0.13333	0.13333	-	-	-
Seq. E	0.14286	0.09524	0.11765	0.13333	-	-
Seq. F	0.09091	0.09524	0.11765	0.13333	0.09524	-

Given the similarity measure (see Table 1.2 for example), for every third sequence C, an additive distance measure is approximated where d(A, B) = d(A, C) + d(C, B), assuming that A, B and C are all related [2]. [1] estimate the additive distance measure between sequences using the equation:

$$d_{\text{kmer}} = 1 - F. \quad (2)$$

TABLE 1.3
Example output of k-mer distance matrix D1

Sequence name	Sequences					
	Seq. A	Seq. B	Seq. C	Seq. D	Seq. E	Seq. F
Seq. A	-		-	-	-	-
Seq. B	0.90476	-	-	-	-	-
Seq. C	0.88235	0.88235	-	-	-	-
Seq. D	0.86667	0.86667	0.86667	-	-	-
Seq. E	**0.85714**	0.90476	0.88235	0.86667	-	-
Seq. F	0.90909	0.90476	0.88235	0.86667	0.90476	-

1.2 Matrix D1 is clustered by UPGMA (Unweighted Pair Group Method with Arithmetic Mean), producing guide tree TREE1.

Given a distance matrix, [1] constructs a guide tree using the UPGMA [2]. The guide tree is based on absolute distances between sequences. This is used to guide the Multiple Sequence Alignment. Using UPGMA, clusters are generated as follows: consider two clusters (subtrees)/sequences L and R to be merged into a new cluster P, which becomes the parent of L and R in the guide tree. Average linkage assigns this distance to a third cluster C:

$$d^{\text{Avg}}{}_{PC} = (d_{LC} + d_{RC})/2. \quad (3)$$

Fig. 2.1 shows an example of generated TREE1 using UPGMA.

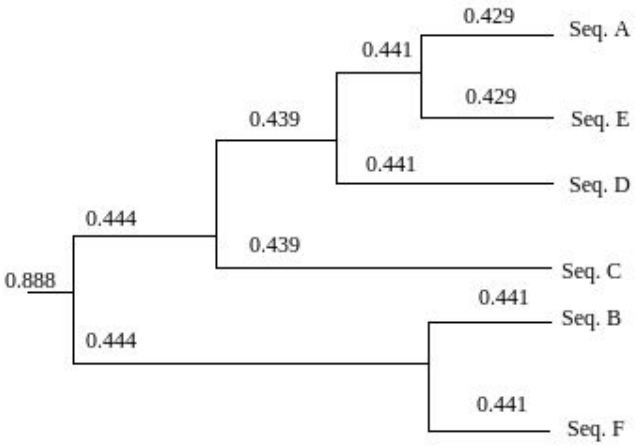

Fig. 2.1: Example TREE1, generated using UPGMA from matrix D1

1.3 A progressive alignment is constructed by following the branching order of TREE1.

A profile is built from an input sequence on any leaf. In prefix order, nodes in the tree are visited (children before their parent). A pairwise alignment of the two child profiles is constructed at each internal node, giving a new profile that is assigned to that node. This induces at the root, a multiple alignment of all input sequences, MSA1. Note that clustering is done at the pair of profile-sequence/profile-profile/sequence-sequence with minimal distance.

Without disrupting the orientation of any specific profile, 2 profiles may be matched. We may insert an indel or an alphabet into any same place of the sequences in the profile to align these profiles without disrupting their internal alignments.

Using the guide tree TREE1, an example order of alignment is determined as follows:

Seq. A and Seq. E alignment:
ATGGTGTGAGAACTTTTGTATAACTTC
GTGGTGTGCCAACTTTACGA-AACTTC

Profile AE and Seq. D alignment:
ATGGTGTGAGAACTTTTGTATAACTTC
GTGGTGTGCCAACTTTACGA-AACTTC

CTGGTGTGCAAACTTTGAGT-------

Seq. B and Seq. F:
GTGGTGTAGGAACTTTGGTCTGG-TGT
GTGGTGTTAGAACTTTTAACCGGCCGT--

Profile AED and Seq. C:
ATGGTGTGAGAACTTTTGTATAACTTC
GTGGTGTGCCAACTTTACGA-AACTTC
CTGGTGTGCAAACTTTGAGT-------

TTGGTGTAATAACTTTTTGCGT-----

Profile AEDC and Profile BF (Multiple Sequence Alignment MSA1 in Stage 1):
ATGGTGTGAGAACTTTTGTATAACTTC
GTGGTGTGCCAACTTTACGA-AACTTC
CTGGTGTGCAAACTTTGAGT-------
TTGGTGTAATAACTTTTTGCGT-----

GTGGTGTAGGAACTTTGGTCTGG-TGT
GTGGTGTTAGAACTTTTAACCGGCCGT

Stage 2, Improved progressive. The approximate kmer distance estimation, which results in a suboptimal tree, is the primary source of error in the draft progressive stage. MUSCLE re-estimates the tree using the kimura distance, which is more accurate but requires an alignment [3].

2.1 The Kimura distance for each pair of input sequences is computed from MSA1, giving distance matrix D2.

Computing the distance in step 2.1 will require the computation of fractional identity D for each pair of sequences. D is computed from a global alignment of two sequences. Below shows an example of the computation.

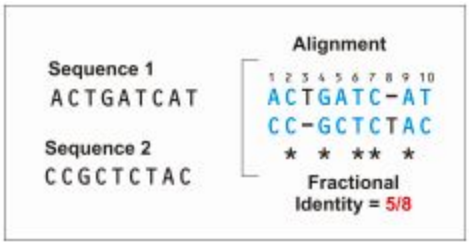

Fig. 2.2: Fractional Identity Example Calculation Between Two Sequences.
Source: Adapted from [10]

To correct the sequences divergence which increases the probability of multiple mutation at a single site, the equation below is used [2]:

$$d_{\text{Kimura}} = -\log_e (1 - D - D^2/5) \qquad (4)$$

Below shows an example fractional identity and kimura distance matrix D2 from MSA1.
Seq. A:ATGGTGTGAGAACTTTTGTATAACTTC
Seq. E: GTGGTGTGCCAACTTTACGA-AACTTC
Seq. D: TGGTGTGCAAACTTTGAGT--------
Seq. C: TTGGTGTAATAACTTTTTGCGT-----
Seq. B: GTGGTGTAGGAACTTTGGTCTGG-TGT
Seq. F: GTGGTGTTAGAACTTTTAACCGGCCGT

TABLE 1.4
Example output of fractional identity between two sequences

Sequence name	Sequences					
	Seq. A	Seq. B	Seq. C	Seq. D	Seq. E	Seq. F
Seq. A	-		-	-	-	-
Seq. B	0.65	-	-	-	-	-
Seq. C	0.64	0.39	-	-	-	-
Seq. D	0.37	0.39	0.79	-	-	-
Seq. E	0.77	0.56	0.62	0.33	-	-
Seq. F	0.59	0.73	0.68	0.26	054	-

TABLE 1.5
Example output of kimura distance matrix D2

Sequence name	Sequences					
	Seq. A	Seq. B	Seq. C	Seq. D	Seq. E	Seq. F
Seq. A	-		-	-	-	-
Seq. B	0.744	-	-	-	-	-
Seq. C	0.791	0.590	-	-	-	-
Seq. D	1.986	1.841	**0.417**	-	-	-
Seq. E	0.458	1.026	0.841	2.276	-	-
Seq. F	0.920	0.550	0.671	3.083	1.102	-

2.2 Matrix D2 is clustered by UPGMA, producing a guide tree TREE2.

Similar process, as in Step 1.2, is applied to D2 to produce TREE2. Fig. 2.3 shows an example output.

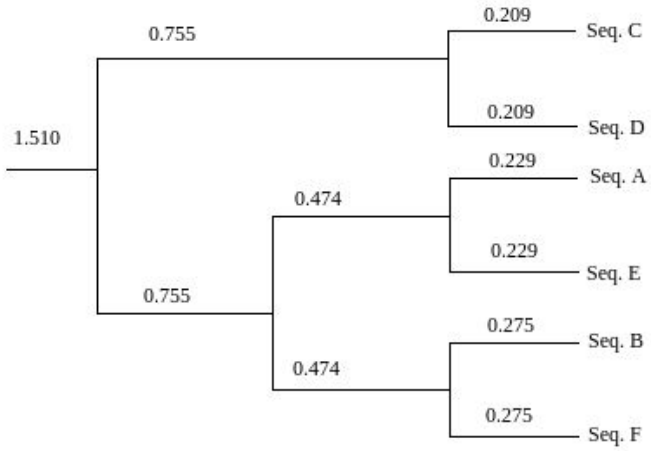

Fig. 2.3: Example TREE2, generated using UPGMA from matrix D2.

2.3 A progressive alignment is produced following TREE2 (similar to 1.3), producing multiple alignment MSA2. This is optimized by computing alignments only for subtrees whose branching orders changed relative to TREE1.

Using the guide tree TREE1, an example order of alignment is determined as follows:

Seq. B and Seq. F alignment (retained since the branch does not changed):

GTGGTGTAGGAACTTTGGTCTGG-TGT
GTGGTGTTAGAACTTTTAACCGGCCGT

Seq. A and Seq. E alignment (retained since the branch does not changed):

ATGGTGTGAGAACTTTTGTATAACTTC
GTGGTGTGCCAACTTTACGA-AACTTC

Seq. C and Seq. D:

TTGGTGTAATAACTTTTTGCGT
CTGGTGTGCAAAC--TTTGAGT

Profile CD and Profile AE:

TTGGTGTAATAACTTTTTGCGT-------
CTGGTGTGCAAAC--TTTGAGT-------

ATGGTGTGAGAAC--TTTTGTATAACTTC
GTGGTGTGCCAAC--TTTACGA-AACTTC

Profile CDAE and Profile BF (Multiple Sequence Alignment MSA2 in Stage 2):

TTGGTGTAATAACTTTTTGCGT-------
CTGGTGTGCAAAC--TTTGAGT-------
ATGGTGTGAGAAC--TTTTGTATAACTTC
GTGGTGTGCCAAC--TTTACGA-AACTTC

GTGGTGTAGGAAC--TTTGGTCTGG-TGT
GTGGTGTTAGAAC--TTTTAACCGGCCGT

Stage 3, Refinement. The third stage performs iterative refinement using a variant of tree-dependent restricted partitioning [11].

3.1 An edge is chosen from TREE2 (edges are visited in order of decreasing distance from the root). Below shows an example of choosing an edge.

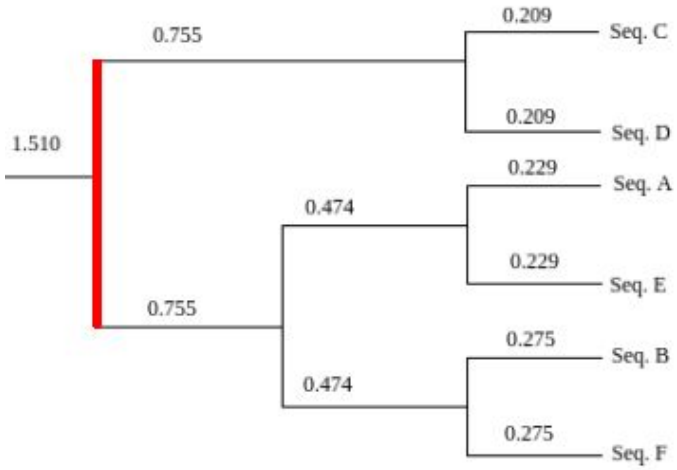

Fig. 2.4: Example TREE2 edge chosen. The color red represents the chosen edge. In this case edge ECFD-AB is chosen.

3.2 TREE2 is divided into two subtrees by deleting the edge. The profile of the multiple alignment in each subtree is computed. Columns which are indels only are discarded. Below shows an example of bipartitioning and profiling.

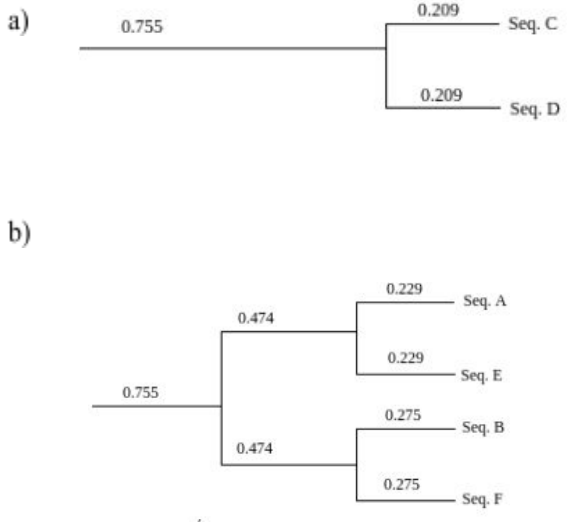

Fig. 2.5: Example TREE2 subtrees.

Using the subtrees in Fig. 1.4, the following profiles are extracted:

Profile CD:
TTGGTGTAATAACTTTTTGCGT
CTGGTGTGCAAAC--TTTGAGT

Profile AEBF:
ATGGTGTGAGAACTTTTGTATAACTTC
GTGGTGTGCCAACTTTACGA-AACTTC
GTGGTGTAGGAACTTTGGTCTGG-TGT
GTGGTGTTAGAACTTTTAACCGGCCGT

3.3 A new multiple alignment is produced by re-aligning the two profiles. Below shows an example re-alignment of two profiles:

TTGGTGTAATAACTTTTTGCGT-------
CTGGTGTGCAAAC--TTTGAGT-------

ATGGTGTGAGAAC--TTTTGTATAACTTC
GTGGTGTGCCAAC--TTTACGA-AACTTC
GTGGTGTAGGAAC--TTTGGTCTGG-TGT
GTGGTGTTAGAAC--TTTTAACCGGCCGT

3.4 If the SP score is improved, the new alignment is kept, otherwise it is discarded.

Steps 3.1- 3.4 are repeated until convergence or until MUSCLE reaches the max iteration defined by the user.

Complete multiple alignments are available at steps 1.3, 2.3 and 3.4, at which points the algorithm may be terminated. We refer to the first two stages alone as MUSCLE-p, which produces MSA2. MUSCLE-p has time complexity O($N^2L + NL^2$) and space complexity O($N^2 + NL + L^2$). Refinement adds an O(N^3L) term to the time complexity.

C. Generating Tree using Maximum-Likelihood Estimation

There are several ways to estimate a phylogenetic tree given the multiple aligned sequenced data. In this study, a phylogenetic tree is estimated using the maximum-likelihood method which was achieved with the help of IQTREE software [1]. But the main idea of this method is to keep generating trees until a tree with the highest likelihood is found, where the likelihood of the tree is the probability of the data given the tree, denoted as L(Tree) = P(Data|Tree) [5]. To simply put, we should find the tree with the highest chance of producing the data.

The algorithm flowchart of IQTREE is shown on figure 3.

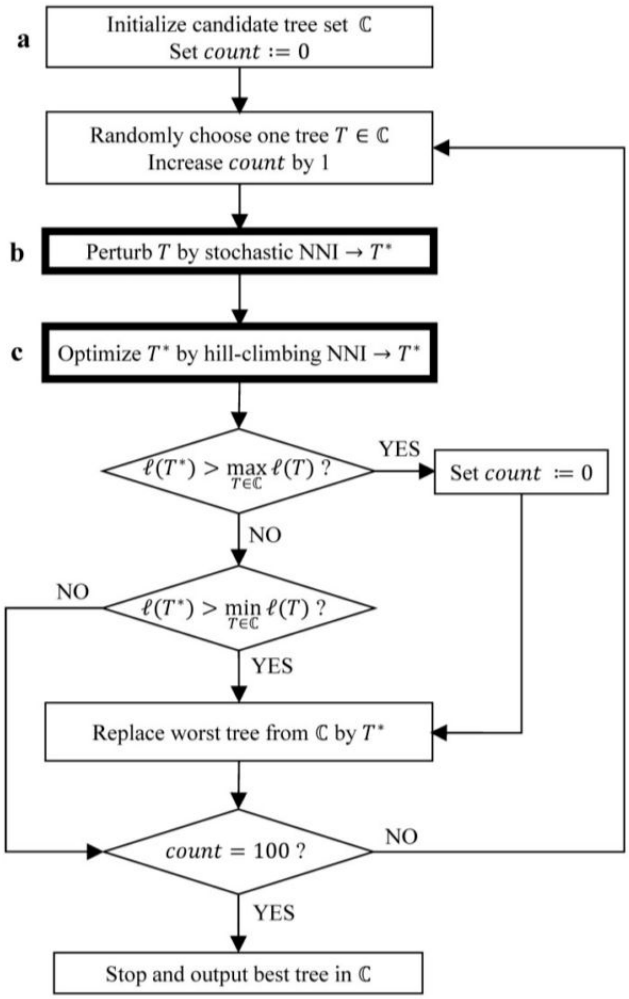

Fig. 3: Internal Algorithm flowchart of IQTREE for estimating a phylogenetic tree.

Source: Adapted from [7]

The algorithm initially starts with 5 trees assigned on variable C. It is generated by estimating 100 trees using Parsimony estimation method. The generated trees will be ranked according to their approximated likelihood, then top 5 trees will be selected for the initial set(Figure 3 box a).

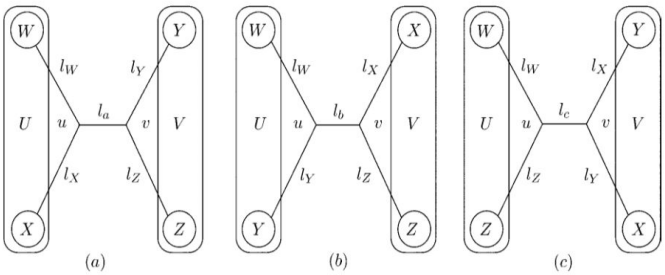

Fig. 4: A sample Nearest Neighbor Interchange (NNI) step applied to an arbitrary inner branch in (a), and two possible results in (b) and (c).

Source: Adapted from [9]

In the next step, a tree is randomly selected from set C, and the count variable will be incremented (the count variable will be used later for terminating the program). The selected tree will be the basis for applying Nearest Neighbor Interchange (NNI) to generate new candidate trees. In the stochastic NNI step (Figure 3 box b), 0.5*(n-3) number of internal branches will be randomly selected, where (n-3) is the number of inner branches in the tree. The selected branches will undergo through NNI step [7]. A sample NNI step is shown on figure 4, where figure 4a is an arbitrary internal edge randomly selected, where U and V indicates the subtrees on the left and right of the internal branch, and the W, X, Y, and Z, are four subtrees under U or V. l_W, l_X, l_Y, and l_Z indicates the lengths of the branches, while l_A, l_B, and l_C are the internal branch lengths of the corresponding selected edge in each sample scenario. An NNI step aims to rearrange the local tree by swapping subtrees. In the case of figure 4a, one scenario is by swapping subtrees X and Y, resulting to X being a subtree of V, and Y being a subtree of U, as shown in figure 4b. Another scenario is shown on figure 4c where subtrees X and Z were interchanged [7][9].

The resulting tree from the NNI step will undergo Branch Length Optimization. It can be achieved by initially calculating the likelihood of the resulting tree, then modify the branch lengths to further maximize the likelihood of the said tree [9]. The general idea for computing the likelihood of the tree is written in the form of L(tree) = P(data|tree, branch lengths, substitution model) [5]. Technically, in IQTREE, the equations below are used to calculate the likelihood of the tree [7].

$$L = \prod_i \sum_{h,h' \in \{A,C,G,T\}} \pi_h L(i = h \mid U) L(i = h' \mid V) P_{hh'}(l). \tag{5}$$

where the π_h is the a priori probability of nucleotide h and $P_{hh'}(l)$ is the probability of nucleotide h to become h' in interval l. Both of the said variables are acquired in the substitution model, which will be explained in later sections. The terms L(i = h|U) and L(i = h'|V) are the conditional likelihood of U and V respectively for any given site *i*, which can be calculated using the equation below [7].

$$L(i = h \mid U) = \left(\sum_{g \in \{A,C,G,T\}} L(i = g \mid W) P_{hg}(l_W) \right) \times \left(\sum_{g \in \{A,C,G,T\}} L(i = g \mid Y) P_{hg}(l_Y) \right), \tag{6}$$

where the value L(i =g|W) is 1 if site i at the tips of W has a nucleotide g, 0 otherwise. Given the equation (5) to calculate the likelihood of the tree, we will adjust the *l* or branch lengths to maximize the L(likelihood). This is a type of optimization problem of one parameter function, and there are several methods to solve this, and Brent's Method[12] is the one being used here [9].

The newly generated tree produced by NNI and Branch Length Optimization steps, either add the new tree in set C or discard it. The said tree is to be added when its likelihood value is higher than the tree with the lowest likelihood on set C, thus discarding the tree with the lowest likelihood in the set C in exchange for the new tree. On the other hand, if the newly generated tree has a lower likelihood value than the tree with the lowest likelihood value in the set, the new tree will be discarded. The algorithm terminates when the counter variable reaches 100, which means that after generating 100 trees using NNI and Branch Length Optimization methods, no more trees produced can be found with a better likelihood value than any trees in set C [7].

D. Models of Sequence Evolution

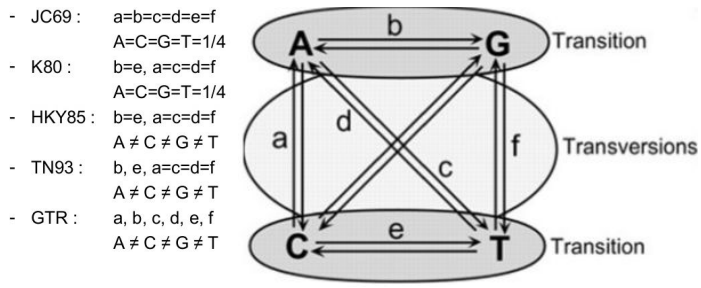

Fig. 4: Substitution Models
Source: Adapted from [15]

Given two aligned DNA sequences, one sequence has an A and one has a C at the third site. Assume that they differ only at the third site and that these sequences originated from the same ancestral sequence. We don't know what ancestral sequence the two sequences have unless we know the ancestral DNA.

Assume that the ancestral sequence had nucleotide A at the third site, there's an infinite number of possibilities to consider. One possibility is that only one mutation has happened which changed A to a C. Another possibility is that the A could have been first mutated to a G before it was mutated again to a C. If we know the actual number of mutations happen to the two sequences since their divergence this mutation count can be used as an evolutionary distance between the two sequences. Since this actual number of mutations is not known, we will use a substitution model to consider all possible mutation sequences.

1) Substitution Model used in [1]

In [1], they used GTR + F + I + G4 to achieve the best predicted model through the tool IQ-TREE [7]. The model Generalised time reversible (GTR) is the most general neutral, independent, and finite sites time reversible model possible in substitution models [6]. GTR allows all six pairs of substitution to have different rates.

Along with GTR, base frequency type +F is defined. This is the default if the model has unequal base frequency.

IQ-TREE supports all common rate heterogeneity across site models. Rate type +I allows for a proportion of invariable sites. Rate type +G alone explains the discrete gamma model with default 4 rate categories.

To sum up, GTR+F+I+G4 explains about the general time reversible model (GTR) with default empirical base frequency (+F) and rate type with invariable site plus discrete gamma model of category number 4 (+I+G4).

E. Bootstrap

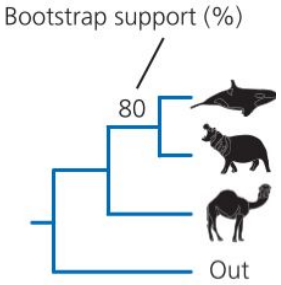

Fig. 5: A sample phylogenetic tree with a defined support value on the branch.
Source: Adapted from [5]

Bootstrapping is a way to find out how much support, sometimes known as the confidence level, we can assume on our generated phylogenetic tree. It can be achieved by generating artificial data sets by random sampling, with replacements or modifications, from the actual data set. By analyzing the generated data, it gives us an insight on the probability of producing the same, if not similar, results when the study is to be replicated many times [5].

The bootstrapping method used by [1] is called 'Ultrafast bootstrap' which consists of three major steps [16]:

1. Initialization Step: Given a multiple sequence alignments with n sequences and m sites, identify the site-pattern frequencies of identical sites and use those frequencies to generate 1000 bootstrapped alignment replicates [1][16].

2. Exploration Step: Perform IQPNNI [17] (just a modified NNI approach which has already been discussed) with the initial tree or the tree to be analyzed. If the likelihood of the resulting tree given any of the

bootstrapped alignment is within the bounds accepted by the algorithm (can be configured by the user), it is added in the set of trees to be used in the consensus in the later steps. Repeat the process until a number of iterations, which is also set by the user, has been reached [16].

3. Summarization Step: Calculate the Bootstrap support by consensus, which simply means by checking the probability on how many times an internal root with the exact subbranches of species has appeared in the bootstrapped tree outputs. Sample is shown on figure 5, and it's done on each internal node in the phylogenetic tree to be analyzed [5][16].

F. BEAST Analysis

1) Temporal Data

Using the sampling dates of the isochronous series, molecular phylogeny can be inferred on a normal time scale of several months. A molecular clock model is used to approximate phylogenies on a natural time scale of months, which is a statistical explanation of the relationship between observable genetic distances and time. To verify whether there is enough time signal available for estimation, a basic regression-based method is used to explore the degree of a time signal in different time series [13]. The equation is as follows as discussed in [13]: for each sequence i let t_i be the sampling time of that sequence, and let $d_{r,i}$ be the genetic distance between that tree tip and the tree root (the so-called 'root-to-tip' distance). If all branches evolve at the same rate, then the phylogenetic timescale can be estimated using the following linear regression model

$$\mathrm{E}[d_{r,i}] = u\,(t_i - t_r) \qquad (7)$$

where u is the rate of sequence evolution and t_r is the time of the tree root. For molecular clock models, regression of root-to-tip genetic distance against sampling time can be used as a diagnostic technique. In [1], the root-to-tip analysis achieved a correlation coefficient of 0.53, which was enough for the Bayesian analysis.

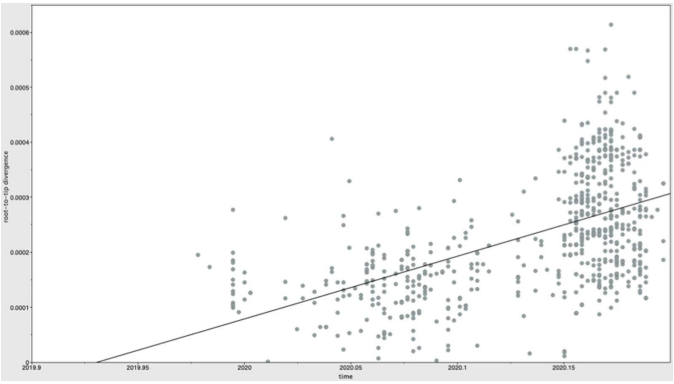

Fig. 6: Root-to-tip analysis of [1].
Source: Adapted from [1]

2) Substitution Model used in [1]

[1] used the HKY nucleotide substitution model. HKY allow stationary frequencies, π, to be unequal, and allow rates of transition and transversion substitutions to differ, κ in the Q matrix [13]. The model can be inferred in [15]

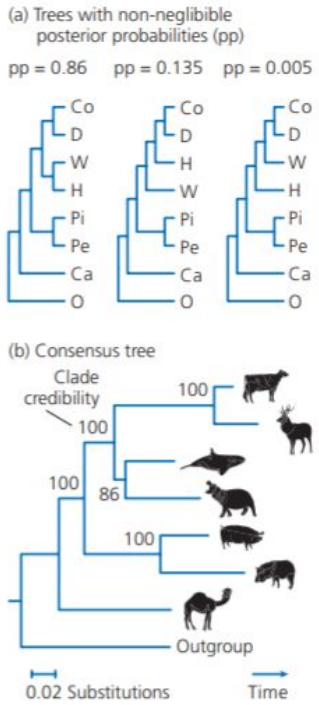

Fig. 7: A sample phylogenetic tree formed using Bayesian Phylogeny Inference
Source: Adapted from [5]

3) Bayesian Phylogeny Inference

When using maximum likelihood as criterion for evaluating trees, probability of the data given the tree is calculated for each possible tree, P(data | tree). This quantity is not the same as the probability of the tree given the data, P(tree | data). And the probability of the tree given the data, is also known as the posterior probability of the tree. Bayes' theorem provides a way to calculate it, by means of this formula. [5]

$$\mathrm{P(tree \mid data)} = \frac{\mathrm{P(data \mid tree)\ P(tree)}}{\mathrm{P(data)}} \qquad \mathbf{(8)}$$

The first term in the numerator is the likelihood of the tree. The second term, P(tree), is the prior probability of the tree. It is the probability assigned to the tree before taking into account the data. The denominator, P(data), is the prior probability of the data. It is the sum of the values we obtain by multiplying each possible tree's likelihood by its prior probability. Because the number of possible trees is usually enormous, it is typically challenging to calculate their posterior probabilities analytically. However, it is possible to find the trees that have non-negligible posterior probabilities,

and estimate what those probabilities are, by using computer software that simulates sampling trees from a population in which each possible tree is represented at a frequency equal to its posterior probability. Clade credibility is computed for each clade as the sum of the posterior probabilities of the trees it occurs in. An example of this can be seen in Figure 7.

V. Paper Extension

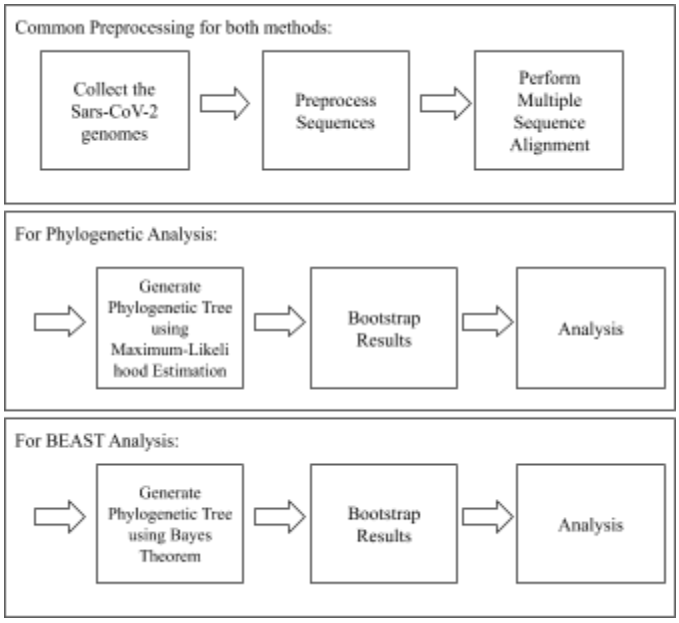

Fig. 8: Paper Extension Pipeline.

A. Common Preprocessing for Both Methods

Collect the Sars-CoV-2 Genomes. Sars-CoV-2 genomes from the Philippines will be collected along with the reported B.1.1.7 variant.

Preprocess Sequences. Sequences with >10% 'N's will be excluded and discrepancies will be checked.

Perform Multiple Sequence Alignment. The preprocessed sequences will be aligned using MUSCLE. The parameters will be the same as mentioned in [1].

B. Phylogenetic Analysis

Generate Phylogenetic Tree using Maximum-Likelihood Estimation. The aligned sequences will be inputted to the IQ-Tree to perform Maximum-Likelihood Estimation. The parameters will be the same as mentioned in [1].

Bootstrap Results. The tree generated will be validated by bootstrap procedure. The parameters will be the same as mentioned in [1].

Analysis. Given the phylogenetic tree, we will analyze how the variant spreads in the Philippines.

C. BEAST Analysis

Generate Phylogenetic Tree using Bayesian Theorem. The aligned sequences will be inputted to the IQ-Tree to perform Maximum-Likelihood Estimation. Then, the output of IQ-Tree will be forwarded to BEAST []. The parameters will be the same as mentioned in [1].

Bootstrap Results. In this stage, the goal is to produce multiple alignment, emphasizing speed over accuracy. This stage builds a progressive alignment.

Analysis.Given the phylogenetic tree, we will analyze how the variant spreads in the Philippines along with the temporal data.

References

[1] B. B. O. Munnink, D. F. Nieuwenhuijse, M. Stein, Á. O'Toole, M. Haverkate, M. Mollers, S. K. Kamga, C. Schapendonk, M. Pronk, P. Lexmond, A. V. D. Linden, T. Bestebroer, I. Chestakova, R. J. Overmars, S. V. Nieuwkoop, R. Molenkamp, A. A. V. D. Eijk, C. Geurtsvankessel, H. Vennema, A. Meijer, A. Rambaut, J. V. Dissel, R. S. Sikkema, A. Timen, and M. Koopmans, "Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands," Nature Medicine, vol. 26, no. 9, pp. 1405–1410, 2020.

[2] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," Nucleic Acids Research, vol. 32, no. 5, pp. 1792–1797, 2004.

[3] R. C. Edgar, BMC Bioinformatics, vol. 5, no. 1, p. 113, 2004.

[4] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," OUP Academic, 10-May-2018. [Online]. Available: https://academic.oup.com/bioinformatics/article/34/18/3094/4994778?login=true. [Accessed: 15-Jan-2021].

[5] J. C. Herron and F. H, "Evolutionary Analysis Fifth Edition," *Illinois: Pearson*. 2014

[6] J. A. H. D. Barry, E. M. P. WM. Brown, M. O. Dayhoff, J. Felsenstein, S. Y. T. Gojobori, K. I. T. Gojobori, W. H. L. T. Gojobori, N. Goldman, H. K. M. Hasegawa, T. G. T. Imanishi, C. R. C. TH. Jukes, M. Kimura, T. M. H. Kishino, G. P. C. Lanave, C.-I. W. W-H. Li, J. L. S. MM. Miyamoto, Y. I. EN. Moriyama, G. A. C. WC. Navidi, J. H. Reeves, M. T. C. K. Ritland, J. F. O. F. Rodriguez, K. Tamura, M. N. K. Tamura, S. Tavare, E. Thompson, W. J. Wilbur, and Z. Yang, "Estimating the pattern of nucleotide substitution," *Journal of Molecular Evolution*, 01-Jan-1987. [Online]. Available: https://link.springer.com/article/10.1007/BF00178256. [Accessed: 16-Jan-2021].

[7] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies," *OUP Academic*, 03-Nov-2014. [Online]. Available: https://academic.oup.com/mbe/article/32/1/268/2925592?login=true. [Accessed: 16-Jan-2021].

[8] Suchard MA;Lemey P;Baele G;Ayres DL;Drummond AJ;Rambaut A; "Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10," *Virus evolution*. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29942656/. [Accessed: 16-Jan-2021].

[9] Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52: 696–704.

[10] A. Manacero, J. M. Machado, L. M. Sato, G. F. D. Zafalon, and E. A. Marucci, "Finding Fractional Identities in Multiple Sequences Using a Fast Parallel Algorithm," *dcce.ibilce.unesp.br*. [Online]. Available: https://www.academia.edu/1180929/Finding_Fractional_Identities_in_M

ultiple_Sequences_Using_a_Fast_Parallel_Algorithm. [Accessed: 17-Jan-2021].

[11] M;, Hirosawa M;Totoki Y;Hoshida M;Ishikawa. “Comprehensive Study on Iterative Algorithms of Multiple Sequence Alignment.” *Computer Applications in the Biosciences : CABIOS*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/7796270/.

[12] Brent, R. 1973. Algorithms for minimization without derivatives. Prentice-Hall, Englewood Clifts, New Jersey.

[13] A. Rambaut, T. T. Lam, L. Max Carvalho, and O. G. Pybus, “Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen),” *OUP Academic*, 09-Apr-2016. [Online]. Available: https://academic.oup.com/ve/article/2/1/vew007/1753488?login=true. [Accessed: 17-Jan-2021].

[14] F. P. G. Aldrich-Blake, W. A. LW. Alvarez, L. W. A. W. Alvarez, A. T. B. S. Anderson, M. H. L. B. S. Anderson, P. Andrews, J. E. C. P. Andrews, B. D. G. CF. Aquadro, A. T. B. BG. Barrell, R. A. E. MJ. Bibb, M. V. S. GG. Brown, M. G. WM. Brown, E. M. P. WM. Brown, W. M. B. RL. Cann, J. L. S. LYE. Chang, A. B. C. RL. Ciochon, R. S. C. RL. Ciochon, E. H. Colbert, N. T. B. JE. Cronin, E. H. W. MH. Day, R. E. Dickerson, J. F. Eisenberg, M. R. Feldesman, J. Felsenstein, A. C. W. SD. Ferris, W. M. B. SD. Ferris, R. D. S. SD. Ferris, K. I. T. Gojobori, M. Goodman, G. B. M. Goodman, B. F. K. M. Goodman, J. C. J. Gribbin, P. A. B. S. Harris, M. Hasegawa, T. Y. M. Hasegawa, T. Y. M. Hasegawa, T. Y. M. Hasegawa, T. D. W. DC. Johanson, T. D. W. DC. Johanson, D. C. W. MJ. Johnson, M. Kimura, T. O. M. Kimura, D. E. Kohne, J. A. C. DE. Kohne, R. L. H. MD. Leakey, R. Lewin, K. A. B. SA. Liebhaber, D. Macdonald, M. C. McKenna, H. H. T. Miyata, J. Moore, U. Nagel, M. Nei, M. J. Novacek, T. S. K. Nozawa, C. E. Oxnard, D. Pilbeam, J. R. Powell, J. J. S. DM. Raup, J. C. B. SM. Raza, J. E. C. VM. Sarich, A. C. W. VM. Sarich, C. E. M. AM. Sarna-Wojcicki, D. E. R. DE. Savage, J. H. Schwartz, P. H. AF. Scott, T. Shotake, J. E. A. CG. Sibley, E. L. Simons, T. U. C. Spolsky, R. L. S. JT. Stern, K. Sugawara, M. K. N. Takahata, M. S. N. Takahata, A. R. Templeton, T. D. White, D. C. J. TD. White, S. S. C. AC. Wilson, O. P. JJ. Yunis, A. L. Zihlman, J. E. C. AL. Zihlman, and L. P. E. Zuckerkandl, “Dating of the human-ape splitting by a molecular clock of mitochondrial DNA,” *Journal of Molecular Evolution*, 01-Jan-1968. [Online]. Available: https://link.springer.com/article/10.1007/BF02101694. [Accessed: 18-Jan-2021].

[15] S. Landtsheer Follow, “Phylogenetics1,” *SlideShare*, 12-Nov-2014. [Online]. Available: https://www.slideshare.net/sebastiendelandtsheer/phylogenetics1. [Accessed: 18-Jan-2021].

[16] Minh BQ, Nguyen MA, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. MolBiolEvol. 30:1188–1195.

[17] Vinh LS, von Haeseler A. 2004. IQPNNI: moving fast through tree space and stopping in time. Mol Biol Evol. 21:1565–1571.