Applicability of phylogenetic network algorithms for representing the evolutionary history of SARS-CoV-2

Rosanne Wallin (2682041)

October 29, 2020

Minor Research Project (XM_0072), 24 EC
Performed at: Centrum Wiskunde & Informatica (CWI), Amsterdam
As part of: Master Bioinformatics and Systems Biology, VU & UvA

Main supervisor: Dr. ir. L.J.J. (Leo) van Iersel
Daily supervisor: Prof. dr. L. (Leen) Stougie
Daily supervisor: Dr. M.E.L. (Mark) Jones
Daily supervisor: Dr. S.M. (Steven) Kelk

Examiner: Dr. ir. K.A. (Anton) Feenstra

Abstract

Phylogenetic networks are used to display complex evolutionary history involving so-called reticulation events, such as genetic recombination. Several methods have been developed to construct such networks, using for example a multiple sequence alignment or multiple phylogenetic trees as input. Coronaviruses are known to recombine frequently, but phylogenetic networks have not yet been used extensively to describe their evolutionary history. Here, we created a workflow to construct phylogenetic networks representing the evolutionary history of SARS-CoV-2 using the TriLoNet, TriL2Net, Tree-Child, Semi-Temporal and Maximum Pseudo-Likelihood algorithms. This included filtering noise from sets of phylogenetic trees by contracting edges based on branch length and bootstrap support, followed by resolving multifurcations. We show that this filtering approach generally reduces the minimum reticulation number and minimum temporal distance of the phylogenetic networks constructed from these trees, while preserving the overall topology. The networks constructed by the TriLoNet, TriL2Net and Tree-Child algorithms show no signs that SARS-CoV-2 itself is a recombinant virus. They do indicate a recombination from an ancestor of the SARS-CoV-2/RaTG13 lineage and the HKU3-1 lineage into the bat-SL-COVZC45 lineage, which was previously described in other research. Our results demonstrate that the TriLoNet, TriL2Net, Tree-Child and Semi-Temporal algorithms are applicable to coronavirus data. However, the constructed phylogenetic networks should be interpreted with care, taking the underlying network constraints and biological plausibility into account. Our workflow may serve as an example for pipelines to preprocess multiple sequence alignments to result in suitable input for phylogenetic network algorithms, providing a base to easily test and validate (multiple) algorithms with different input data and filtering options.

Introduction

Since the beginning of 2020, the entire world has been greatly impacted by the outbreak of COVID-19, caused by the Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2) virus. This pandemic follows on previous outbreaks of Severe acute respiratory syndrome (SARS) in 2003 and Middle East respiratory syndrome (MERS) in 2012, caused by the SARS and MERS coronaviruses (Drosten et al., 2003; Zaki et al., 2012). This series of outbreaks has led to a lot of public and scientific interest in the origin and evolution of these coronaviruses, which could be crucial to prevent possible future outbreaks by other coronaviruses or to accelerate the development of medicines and vaccines.

Biological background of coronaviruses

Coronaviruses (*Orthocoronavirinae*) are positive-sense single-stranded unsegmented RNA viruses with relatively large genomes (26.4 to 31.7 kb) (Woo et al., 2010). They infect various mammal species and can relatively easily switch between hosts, including animal-to-human transmission (zoonosis) (Graham and Baric, 2010). Bats and rodents are known to serve as a natural reservoir for viruses, because they can host one or multiple viruses while showing no or only mild symptoms of disease (Ye et al., 2020). They have been found to host in total 61 (bats) and 68 (rodents) zoonotic viruses, including coronaviruses (Luis et al., 2013). Research into the origin of the SARS coronavirus suggested that although civets have probably served as an intermediate host for transmission to humans, the virus originated from bats (Graham and Baric, 2010; Cui et al., 2019) and the MERS coronavirus was transmitted to humans from dromedary camels, but several MERS-like coronaviruses have also been found in bats (Dawson et al., 2019). Two more human coronaviruses (HCoV-NL63 and HCoV-229E) are likely to have originated from bats, while two others (HCoV-OC43 and NCoV-HKU1) are thought to originate from rodents (Cui et al., 2019).

These natural reservoirs of bats and rodents create opportunities for genetic recombination, which can occur when different virus lineages infect the same host cell. For unsegmented RNA viruses this recombination is thought to be caused by a process called template switching. During replication, RNA poly-

merase can switch from one template sequence to another (while the newly synthesised RNA remains bound to the polymerase) and resume synthesis there. When templates from two different viruses are present in one cell, template switching can result in a new recombinant virus containing genetic material from both viruses (Lai and Cavanagh, 1997; Simon-Loriere and Holmes, 2011). Coronaviruses specifically are known to be capable of highly frequent recombination and the recombination breakpoints seem to be distributed randomly across the genome (Lai and Cavanagh, 1997). A study by Zhang et al. (2005) showed that the SARS-CoV genome contains at least seven regions that potentially result from recombination events involving six other coronaviruses. Another study concludes that SARS-CoV is probably the result of several recombination events in different bat species (Luk et al., 2019).

Regarding the origin of SARS-CoV-2, the closest known relative of SARS-CoV-2 is a bat SARS-related coronavirus with ∼96% overall genetic identity, called RaTG13 (Zhou et al., 2020). However, several studies conclude that these two lineages have already diverged several decades ago, indicating that RaTG13 is not the direct progenitor of SARS-CoV-2 (Boni et al., 2020; Wang et al., 2020). Also, the two genomes show some divergence in the receptor-binding domain (RBD) of the spike protein where SARS-CoV-2 does and RaTG13 does not contain the key residues for binding to the human ACE2 receptor (Andersen et al., 2020). Multiple lineages of coronaviruses have been discovered in Malayan pangolins, with strong similarity to SARS-CoV-2 in the RBD, including the aforementioned key residues (Lam et al., 2020; Andersen et al., 2020). The exact nature of the genetic relationship between pangolin coronaviruses and SARS-CoV-2 is still debated. Li et al. (2020) suggests that recent recombination between a RaTG13-like virus with a pangolin virus resulted in SARS-CoV-2 with the gained ability to bind human ACE2. Other studies argue that a recombination of the RBD region from a yet unknown coronavirus into RaTG13, after divergence from SARS-CoV-2 and the pangolin viruses, is a more plausible explanation for the variation in sequence similarity (Wang et al., 2020; Boni et al., 2020).

Altogether it is clear that the process of recombination should be taken into account when investigating the origin of coronaviruses. Traditionally, evolutionary history is inferred and displayed by constructing phylogenetic trees, but these trees can only represent vertical evolutionary processes, such as the lineal descent from parent to offspring. Recombination is a horizontal evolutionary event, where two lineages recombine into a new lineage, which cannot be depicted in a tree. To represent such more complex evolutionary relationships, phylogenetic networks can be used. Different network types can be distinguished, which serve different purposes. Here we will limit ourselves to explicit rooted phylogenetic networks, which aim to represent evolutionary history rather than to solely display patterns in the data.

Phylogenetic networks

In phylogenetic networks, events like recombination, but also hybridisation (e.g. in yeast) and horizontal gene transfer (e.g. in bacteria) are displayed by so-called reticulations. Such a reticulation node has more than one ingoing edge (Figure 1). An important measure for the complexity of a network is its reticulation number, which is the number of additional edges in the network compared to a tree structure (Figure 1). For a formal definition, see e.g. van Iersel and Moulton (2014).

Several methods have been developed to construct phylogenetic networks, using various types of input data and approaches and having different limitations in terms of input size and complexity. The input data is usually (derived from) a multiple sequence alignment (MSA) or a set of trees constructed for different parts of the genome. Some methods require all trees to be binary (every parent node has exactly two child nodes), others can handle non-binary trees as well. Non-binary trees contain so-called multifurcations, where parent nodes have more than two child nodes. These are typically used to denote uncertainty about the exact sequence of bifurcation events occurring at that point.

Phylogenetic network algorithms often pose constraints on the structure of the network, for example to reduce the mathematical complexity of construct-

ing the network and to avoid overfitting. One example of this is limiting the level of the network, which is the maximum number of reticulations in each reticulate part of the network (Figure 1). For a formal definition, see e.g. van Iersel and Kelk (2011). TriLoNet is a method that constructs level-1 networks from an MSA. It computes a small level-1 network for each group of three taxa, called a trinet and then combines these trinets to construct a corresponding network. It has been used to construct networks for HIV and hepatitis B data sets with recombinant sequences (Oldman et al., 2016). TriL2Net is a similar algorithm, which can construct networks up to level 2 from trinets (Kole, 2020).

Another type of network is a tree-child network, in which every parent node has at least one child node that is not a reticulation (Figure 1). The Tree-Child algorithm (van Iersel et al., 2019) uses multiple binary trees as input to construct a tree-child network which displays all input trees with a minimum reticulation number. It has been shown to be able to construct networks as complex as level-11 for synthetic data and level-21 for data from bacterial and archaeal genomes (van Iersel et al., 2019).

An even more restricted type of network is a temporal network, which is a tree-child network in which it is possible to assign each node a timestamp, such that every tree node has a more recent timestamp than its parent node and any reticulation node has the same timestamp as all of its parent nodes (Figure 1). The Temporal algorithm finds a temporal network from multiple binary trees (if it exists) (Borst et al., 2020). An extension of this algorithm is the Semi-Temporal algorithm, which finds a tree-child network solution that deviates as little as possible from the temporal restriction (Borst et al., 2020). It has been shown to be often faster but sometimes much slower compared to the Tree-Child Networks algorithm, depending on the input data (Borst et al., 2020).

Next to such combinatorial methods, other methods exist that use a heuristic search approach, which is often computationally less expensive. The PhyloNet package (Than et al., 2008; Wen et al., 2018) contains several such methods, which also take incomplete lineage sorting (ILS) into account. ILS is a phenomenon that causes trees constructed for a specific part of the genome to differ from the overall species tree, while no exchange of genetic material has occurred (Maddison, 1997). The maximum pseudo-likelihood (MPL) algorithm in this package uses multiple trees as input and calculates the frequencies of so-called triples (trees with three leaves) in the input trees. It searches for possible networks and calculates the pseudo-likelihood of those networks using the triple frequencies. One disadvantage of this method is that it does not guarantee finding the optimal network and sometimes multiple optimal solutions exist. However, this algorithm can handle larger data sets than the related maximum likelihood algorithm and has shown good results in inferring hybridization in a yeast data set (Yu and Nakhleh, 2015).

Research goal

Currently, the evolutionary history of coronaviruses is often represented by (multiple) phylogenetic trees, but phylogenetic networks may be more suitable to represent their complex histories with frequent recombination. The aforementioned phylogenetic network methods have previously been tested on synthetic and biological data sets, but their suitability for coronavirus data has not been studied before.

Here, we explore the applicability of phylogenetic network methods to infer and display the evolutionary history of coronaviruses. To this end, we create a workflow to test five different algorithms for constructing phylogenetic networks (TriLoNet, TriL2Net, Tree-Child, Semi-Temporal and MPL) with various inputs derived from an MSA of SARS-related viruses. This includes a filtering approach to reduce noise in the input trees for the tree-based algorithms. Using this workflow, we aim to answer the following three research questions:
1. What are the limitations of these algorithms for coronavirus data in input size and complexity?
2. What is the influence of our filtering approach on the complexity and topology of the phylogenetic networks constructed by the algorithms?
3. How do the phylogenetic networks constructed by these algorithms relate to the evolutionary history of SARS-CoV-2 as known from the literature?

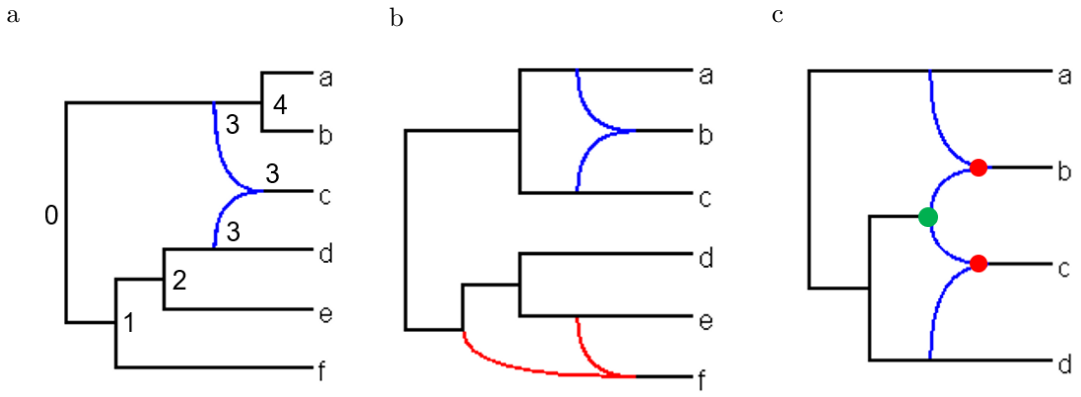


Figure 1: Three examples of phylogenetic networks with reticulations indicated in red and blue. a) Temporal network (labelled) with reticulation number 1 and level 1. b) Tree-Child non-temporal network with two reticulate parts, both containing one reticulation. It has reticulation number 2 and level 1. c) Network that is not tree-child, because both child nodes (in red) of the green parent node are reticulations. It has reticulation number 2 and level 2.

Methods

Workflow design

The workflow was implemented in three main scripts, as depicted in Figure 2. The first script builds binary phylogenetic trees from an MSA for different parts of the genome, as indicated by a set of breakpoint locations. Optionally, it removes specified taxa from the MSA. The second script runs sequence-based phylogenetic network methods from an MSA. The third script runs tree-based phylogenetic network methods given a set of trees, optionally after filtering. The details of the input data, the preprocessing and filtering steps and the different phylogenetic network algorithms will be discussed in the next subsections.

We included some additional steps to convert files between formats, as the methods are limited in the input formats they can handle. We used the sequential PHYLIP (Felsenstein, 2020b) and NEXUS formats (Maddison et al., 1997) for MSAs and the Newick Standard (Felsenstein, 2020a) and NEXUS formats (Maddison et al., 1997) for trees. Some algorithms require specific versions of these formats, e.g. Newick Standard without branch lengths and bootstrap support values (SI Spreadsheet 1).

Input data

We used a data set from Grimm and Morrison (2020), which has been derived from roughly 300 genomes of SARS-like viruses. As input for the workflow we used the MSA that includes a selection of 21 of these taxa and has been adjusted to exclude for example badly aligned regions (Grimm and Morrison, 2020). After inspection of the MSA, we corrected the location of a gap in the start codon of *ORF3a* (from A-TG to -ATG) for a few taxa.

During early stages of the research, we saw that none of the tree-based phylogenetic network methods was able to return a network for these 21 taxa (within a reasonable time limit). Excluding an increasing number of taxa from the trees showed that for a selection of 12 taxa, some of the algorithms did return a network. Therefore, we originally chose to use three selections (A, B and C) containing respectively 12, 9 and 7 taxa (Table 1). In the taxon selections we aimed to include a diverse set of SARS-related coronaviruses, based on their pairwise distances (Grimm and Morrison, 2020), to represent as much of this subgenus as possible. The exclusion of multiple closely related taxa can also prevent unnecessary reticulations caused by disagree-

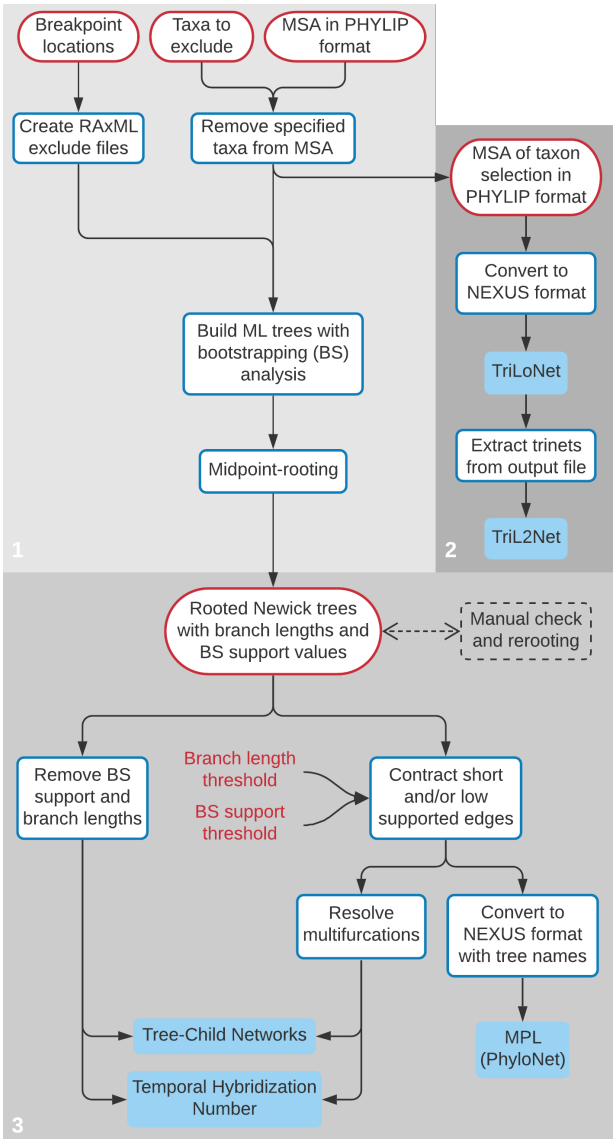


Figure 2: Flowchart depicting the main steps in the workflow, as implemented in the main scripts (1, 2 and 3). Colours indicate input files (boxes with red outline), input parameters (red text), preprocessing and filtering steps (boxes with blue outline) and phylogenetic network algorithms (blue boxes). For clarity, intermediate data files and algorithm output files are not included.

ment between trees. In the smaller taxon selections, the focus lies on taxa that seemed related to SARS-CoV-2, based on trees and networks constructed with selection A. After analysing the networks resulting from these original selections, we included additional taxon selections excluding bat-SL-CoVZC45 (A-, B- and C-), which are otherwise identical to their corresponding original selection (A, B and C). This choice is further explained in the Results section.

Accession	**Isolate name**	**Host**	**Sel.**
MN908947	Wuhan-Hu-1	Human	ABC
MN996532	RaTG13	Bat	ABC
MT121216	MP789	Pangolin	ABC
MT040334	PCoV_GX-P1E	Pangolin	ABC
MG772933	Bat-SL-CoVZC45	Bat	ABC
NC_004718	Tor2	Human	ABC
DQ022305	HKU3-1	Bat	ABC
DQ412042	Rf1	Bat	AB
DQ412043	Rm1	Bat	AB
JX993988	Cp/Yunnan2011	Bat	A
KP886808	YNLF_31C	Bat	A
KJ473816	BtRs-YN2013	Bat	A

Table 1: GenBank genome accession IDs, isolate names and hosts of the selected SARS-related viruses. The last column states in which of the selections A (n=12), B (n=9) and C (n=7) they were included. Note that Wuhan-Hu-1 and Tor2 are the reference genomes for SARS-CoV-2 and the SARS coronavirus respectively.

We used two sets of breakpoint locations to divide the genome into different parts. One set consists of 9 equally-sized blocks covering the entire genome (as suggested by Grimm and Morrison (2020)), which is an arbitrary way of dividing the genome without utilising any biological information or information from the genomic data itself. The other set consists of the genes that make up the SARS-CoV-2 genome. We transferred the gene locations annotated for SARS-CoV-2 in the NCBI GenBank (Benson et al., 2012) to the MSA, taking into account the gaps inserted in the SARS-CoV-2 genome during alignment and the deleted regions as annotated by Grimm and Morrison (2020). The *ORF1ab*/*ORF1a* gene

covers more than half of the genome and contains several non-structural proteins (nsp's), which were therefore included as separate 'genes'. An overview of the genes, their original lengths and the lengths of the corresponding sequences in the MSA is given in SI Table 1. *NSP11* and *ORF8* were excluded from the gene set, because of their extremely short sequence lengths in the MSA. The BtRs-YN2013 sequence did not contain the *ORF10* gene, so this gene was excluded for taxon selections A and A-.

Constructing trees

Binary trees were built under the maximum-likelihood (ML) criterion in RAxML (Stamatakis, 2014) using the function for rapid bootstrap (BS) analysis with 100 replications (Stamatakis et al., 2008), followed by a search for the best-scoring ML tree (with the GTR-CAT model). From the RAxML output we used the best-scoring ML trees, including branch lengths and BS support values [1].

We rooted the trees using midpoint-rooting in Dendroscope (Huson and Scornavacca, 2012) and visually inspected all trees. We rerooted trees manually, if necessary, to ensure that all trees were rooted consistently. In some blocks and genes, the sequences from multiple taxa were identical, so RAxML assigned extremely small branch lengths, which Dendroscope automatically converted to the scientific notation (e.g. 1E-6). Because this format is not suitable for all phylogenetic network methods, we replaced these branch lengths with a value of 0.0.

Filtering trees

To filter out noise from the trees, we performed edge contraction in Dendroscope (Huson and Scornavacca, 2012) based on branch length and BS support values, followed by resolving multifurcations. Short edges were contracted, so that splits that are based on a very small amount of variation are removed. This also ensures that multiple taxa with identical sequences are represented as one multifurcation. We also contracted edges based on their BS support values, to filter out low confidence clades. We chose one threshold for branch length (l=0.01) and two thresholds for BS support value (s=70 and s=90), based on the observed branch lengths and BS support values in the block-based trees built for taxon selection A. We also looked at the values of the gene-based trees (selection A) to confirm the suitability of these thresholds for the gene-based experiments. To observe the influence of these thresholds, we used all possible combinations of the thresholds, including the option of no threshold, resulting in six different options for edge contraction.

To convert the non-binary trees that resulted from the edge contraction to binary trees, we used the Data Transformer for Real-World Data (an additional feature of the Tree-Child Network method (van Iersel et al., 2019)). This algorithm tries to resolve the multifurcations in sets of non-binary trees in such a way that there is as much consistency between the resulting binary trees as possible. It outputs all unique trees without branch lengths and BS support values, but it also introduces quotation marks into the taxon labels. We added a step to remove these quotation marks, which is necessary for the trees to be suitable for the Semi-Temporal algorithm.

The MPL method from PhyloNet can handle non-binary trees and makes use of the branch lengths (unlike the Tree-Child Network and Semi-Temporal algorithms). Therefore, we used the edge contracted trees with branch lengths and BS support values directly as input for MPL, without resolving the multifurcations.

Phylogenetic network methods

We started by investigating 11 methods for constructing phylogenetic networks (SI Spreadsheet 1), but not all of them were included in the final workflow. This was mostly because of practical aspects: they could not be run from a command-line interface, they required too much additional preprocessing of input

[1] Bootstrapping is a method in which replicate matrices are created by randomly selecting (with replacement) columns from the original matrix. In this case, the original matrix is the MSA and the columns consist of nucleotides. For each of these replicate matrices a tree is built. BS support values are then calculated for all clades as the percentage of BS trees in which that clade appears, thereby indicating the confidence that the clade is correct (Efron et al., 1996; Felsenstein, 1985).

data or they did not function properly.

From the sequence-based methods we included TriLoNet (Oldman et al., 2016) and TriL2Net (Kole, 2020). For these methods the only variations in input were the six taxon selections. TriL2Net requires trinets as input, so the trinets constructed by TriLoNet were extracted and used as input for TriL2Net. Note that these trinets are maximum level-1 and consequently the networks constructed by TriL2Net are maximum level-1 as well.

As tree-based methods we included the Tree-Child (van Iersel et al., 2019), Semi-Temporal (Borst et al., 2020) and Maximum Pseudo-Likelihood (MPL) algorithms (maximum of 10 reticulations, maximum of 1000 network topologies to examine, 10 processors) (Yu and Nakhleh, 2015). These were run with every possible combination of the two breakpoint sets, the six taxon selections and the six edge contraction options, resulting in 36 instances for each method. The MPL algorithm returns five networks (with the highest log likelihood) for each input, while all other algorithms return one optimal network for each input.

Our earlier experiments with this data set showed that TriLoNet and TriL2Net always returned a network within a few minutes and that the MPL method always returned networks within 30 minutes. Therefore, we did not put a limit on the runtime for these algorithms. The Tree-Child Network and Semi-Temporal algorithms did not always find a solution within 30 minutes, so we limited their runtimes to 5 minutes. Their computational times increase exponentially with the reticulation number, so a longer time limit would only marginally increase the input sizes for which a solution can be returned.

Results

TriLoNet and TriL2Net

Both TriLoNet and TriL2Net were able to find a network for each of the taxon selections and even for the original selection of 21 taxa (not included in the workflow). Two main clades appear in all of the resulting networks, which we will refer to as the SARS-CoV-2 clade (including Wuhan-Hu-1, RaTG13, MP789 and PCoV_GX-P1E) and the SARS clade (including Tor2, HKU3-1, Rf1, Rm1, Cp/Yunnan2011, YNLF_31C and BtRs-YN2013). The networks from TriLoNet and TriL2Net for taxon selections A, B and C all consistently show one reticulation from an ancestor of Wuhan-Hu-1 (SARS-CoV-2) and RaTG13 and HKU3-1 into bat-SL-CoVZC45 (Figure 3).

Because of the level-1 restriction of these methods, it might be that not all reticulation signals in the data are represented in the constructed networks. Hence, we used taxon selections without bat-SL-CoVZC45 as a possibility to find other reticulations as well. For two of these selections (A- and B-) the networks all show one reticulation within the SARS clade (Figure 3). The locations of the ingoing and outgoing edges of this reticulation and the involved taxa differ between the two algorithms and between the two taxon selections. The networks resulting from the smallest selection without bat-SL-CoVZC45 (C-) both show no reticulations at all.

The topology of the SARS-CoV-2 clade is identical in all networks, but the positioning of the taxa in the SARS clade is inconsistent between the two algorithms as well as between taxon selections. This can be seen when comparing the networks in Figure 3.

Input trees

All combinations of breakpoints, taxon selections and edge contraction thresholds resulted in 36 sets of trees to use as input for the tree-based phylogenetic network algorithms. The number of unique trees in each of these tree sets is an indication of the complexity of this input data. Excluding bat-SL-CoVZC45 from taxon selections B and C drastically decreased the number of unique input trees, for example from 20 (selection B) to 12 (selection B-) for the gene-based tree set without filtering (SI Table 2). Also, the number of unique trees often slightly decreased with stricter edge contraction thresholds, but in four cases it increased. The tree sets with the smallest taxon selection (C-) consist of only one unique tree, regardless of whether blocks or genes were used as breakpoints.

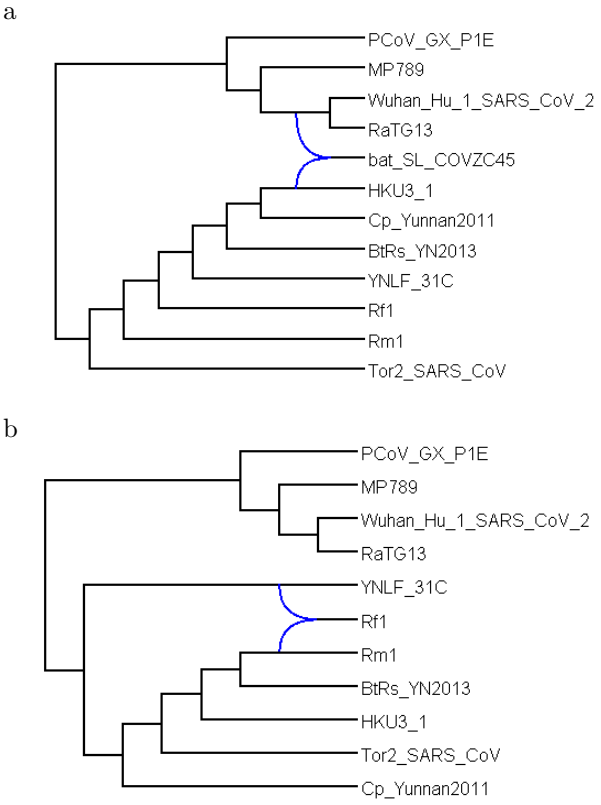


Figure 3: Networks constructed by TriLoNet based on the MSA with taxon selection A (fig. a) and A- (fig. b). Both networks have reticulation number 1.

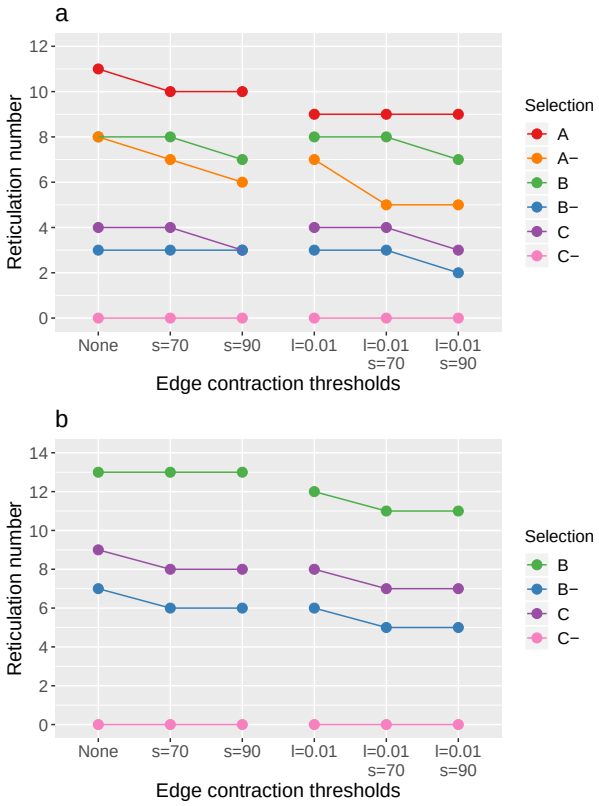


Figure 4: Reticulation numbers of the networks constructed by the Tree-Child algorithm from the block-based (a) and gene-based (b) input trees, depending on the combination of branch length (l) and BS support (s) thresholds. The six taxon selections with (A, B and C) and without bat-SL-CoVZC45 (A-, B- and C-) are indicated by the different colours. No networks could be constructed for the gene-based trees with taxon selections A and A-.

Tree-Child algorithm

The Tree-Child algorithm was able to solve input sets with up to 12 taxa for the block-based trees and up to 9 taxa for the gene-based trees, but it did not return a network for taxon selections A and A- when using the genes as breakpoints. Reticulation numbers are higher for gene-based trees compared to block-based trees, indicating that the former contain more topological differences, resulting in a more complex network. The highest reticulation number (and level) for which a network could be returned was 13. Taxon selections without bat-SL-CoVZC45 showed a large decrease in reticulation number compared to their corresponding original selection, indicating that this taxon is involved in many reticulations (Figure 4).

Regarding the influence of filtering, the reticulation number of the network produced by the Tree-Child algorithm generally decreased (and never increased) after performing edge contraction followed by resolving multifurcations (Figure 4). Compared to no edge contraction at all, the branch length threshold (l=0.01) reduced the reticulation number for five inputs and the BS support thresholds (s=70 and s=90) reduced the reticulation number for respec-

tively four and six inputs. The strictest threshold combination for edge contraction (l=0.01 and s=90) always resulted in the lowest reticulation number (although not exclusively).

With regard to the topology of the networks constructed by the Tree-Child algorithm, it stood out that the reticulations frequently have more than two (and up to five) incoming edges (Figure 5). All constructed networks consistently show the previously mentioned SARS-clade and SARS-CoV-2 clade and a reticulation from these clades into bat-SL-CoVZC45 (e.g. Figure 5). Additional reticulations occur mostly in the SARS-clade and vary between block-based and gene-based networks and between taxon selections. Sometimes additional reticulations are shown in the SARS-CoV-2 clade, but for the selections without bat-SL-CoVZC45, the SARS-CoV-2 clade is strictly tree-like in all networks.

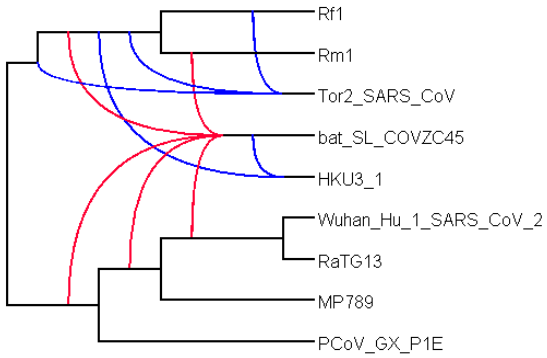


Figure 5: Network constructed by the Tree-Child algorithm from the block-based trees consisting of taxon selection B, after edge contraction (l=0.01 and s=90) followed by resolving multifurcations. The network has reticulation number 7 and level 7.

Semi-Temporal algorithm

The Semi-Temporal algorithm found a solution for almost all instances of the block-based trees. For taxon selections A, A- and B no solution was found if no edge contraction was performed, but a solution was found for all other edge contraction options (Figure 2). With the gene-based trees as input, no solution could be found for the larger taxon selections (A, A- and B). For selection C, a solution could only be found if edge contraction was performed with any of the thresholds (Table 2). Altogether the algorithm was able to solve (after filtering) input sets with up to 12 taxa based on the 9 blocks and input sets with up to 8 taxa based on the 24 genes. The solutions have reticulation numbers up to 11 and temporal distances up to 7 (SI Spreadsheet 2).

In general, the temporal distance for which a network was found was lower or equal when edge contraction was performed. However, in one case (block-based trees with taxon selection A) the temporal distance increased from 2 to 3 when increasing the BS support threshold from 70 to 90 (with l=0.01) (Table 2). The reticulation numbers of the solutions found by this method are almost always equal to those found by the Tree-Child algorithm (SI Spreadsheet 2). For a few instances the reticulation number is slightly higher, which is probably because the Semi-Temporal algorithm finds the minimum temporal distance rather than the minimum reticulation number.

Maximum Pseudo-Likelihood

The MPL algorithm was able to construct networks for all inputs. The complexity of the returned networks differs greatly between the various inputs, with reticulation numbers between 0 and 11 (SI Table 3). Sometimes this variation exists within the five networks from one input as well, e.g. in one case three of the networks have no reticulations and the other two have reticulation numbers 6 and 7. Also, the reticulation numbers do not show any consistent relationship with the edge contraction thresholds or the number of included taxa (SI Table 3).

Most of the networks show the previously mentioned division into a SARS and SARS-CoV-2 clade, regardless of the different breakpoint locations, taxon selections and edge contraction options. If the networks show one or more reticulations, they nearly always show a reticulation from these two clades into bat-SL-CoVZC45 (if this taxon was included in the taxon selection) (e.g. Figure 6a). Some of the returned networks contain chains of reticulations where the output node of one reticulation is the input

		Edge contraction thresholds					
		None	None	None	l=0.01	l=0.01	l=0.01
Breakpoints	**Selection**	None	s=70	s=90	None	s=70	s=90
Blocks	A (n=12)	–	3	3	2	2	3
	A- (n=11)	–	4	3	5	3	3
	B (n=9)	–	5	3	5	5	3
	B- (n=8)	2	2	2	2	2	0
	C (n=7)	0	0	0	0	0	0
	C- (n=6)	0	0	0	0	0	0
Genes	A (n=12)	–	–	–	–	–	–
	A- (n=11)	–	–	–	–	–	–
	B (n=9)	–	–	–	–	–	–
	B- (n=8)	7	5	5	6	4	4
	C (n=7)	–	6	6	6	4	4
	C- (n=6)	0	0	0	0	0	0

Table 2: Minimum temporal distance found by the Semi-Temporal algorithm for different breakpoint location sets, taxon selections (n = number of taxa) and thresholds for edge contraction based on branch length (l) and BS support values (s). "–" indicates that no solution was found within the runtime limit of 5 minutes.

node for a next reticulation, as visible in Figure 6b. Surprisingly, for the smallest taxon selection (C-) the MPL algorithm returned networks with reticulations (Figure 6b), while both the block-based and gene-based trees can be displayed by one unique tree for this taxon selection. This may be due to the fact that MPL uses the branch lengths in its calculations, which differ between the trees.

Across all MPL networks, there is some inconsistency in the positions of the taxa in the SARS clade, but also in the SARS-CoV-2 clade. When using the block-based trees with taxon selection B (no edge contraction) as input, this positioning is very atypical in all of the five returned networks. None of them has Wuhan-Hu-1 (SARS-CoV-2) and RaTG13 positioned as siblings (e.g. Figure 6a), even though they are siblings in all corresponding input trees.

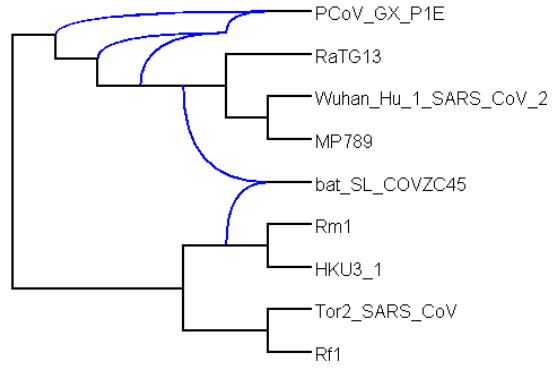
Conclusion and Discussion

We designed and implemented a workflow to test the applicability of the TriLoNet, TriL2Net, Tree-Child, Semi-Temporal and MPL algorithms for representing the evolutionary history of SARS-CoV-2. We analysed the limitations in input size and complexity of these algorithms, the influence of our filtering approach on the constructed networks and the biological interpretation of these networks.

Input size and complexity

The limitations in input size and complexity vary between algorithms. TriLoNet and TriL2Net solved input up to at least 21 taxa for this type of data. The Maximum-Pseudo Likelihood algorithm was able to handle larger inputs (up to 12 taxa in 23 trees) than the Tree-Child algorithm (up to 9 taxa in 24 trees and up to 12 taxa in 9 trees). Previous research showed that the Tree-Child algorithm was able to solve 306 of 630 instances consisting of up to 150 taxa in 8 trees (van Iersel et al., 2019), so the limit found here in terms of input size is notably lower. This might be due to the different biological source of the data (viruses versus bacteria and archaea) resulting in differences in reticulation numbers. We showed that the Tree-Child algorithm was able to construct networks with levels up to 13, which is inbetween the previously shown limits for synthetic (level-11) and biological (level-21) data (van Iersel et al., 2019). The

a



b

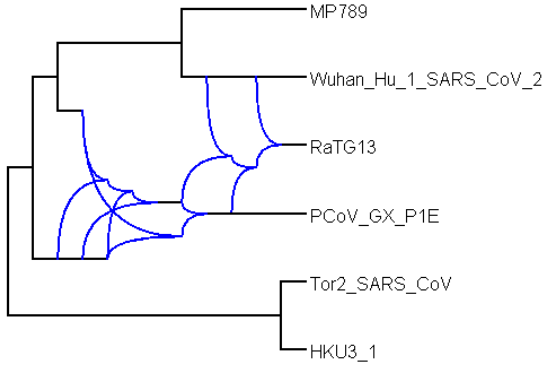


Figure 6: Two of the networks constructed with the maximum pseudo-likelihood method in PhyloNet. The network in fig. a was constructed from the block-based trees with taxon selection B (no edge contraction) and has reticulation number 3 and level 3. The network in fig. b was constructed from the gene-based trees with taxon selection C- (no edge contraction) and has reticulation number 8 and level 8.

Semi-Temporal algorithm was slightly more limited by input size (up to 8 taxa in 24 trees) and complexity (reticulation numbers up to 11 and temporal distances up to 7) compared to the Tree-Child algorithm. However, the Semi-Temporal algorithm could be useful in cases where networks with minimum temporal distance instead of minimum reticulation number are biologically more relevant.

Influence of filtering

Filtering noise by contracting edges (based on branch length and bootstrap support) followed by resolving multifurcations reduced the reticulation number of the networks constructed by the Tree-Child algorithm, while the general topology remained the same. It also, with one exception, reduced the temporal distance of the networks constructed by the Semi-Temporal algorithm. For some inputs, this algorithm was able to construct a network within the runtime limit only after filtering. However, this filtering approach did not consistently reduce the reticulation numbers of networks constructed by the MPL algorithm.

Even after filtering, the networks constructed by the Tree-Child algorithm still contained reticulations with many incoming edges. This topology is biologically not very likely, so it may indicate that the input trees still contain some noise. Stricter thresholds for edge contraction could possibly be used to reduce more noise, but this can also lead to loss of relevant information. Additionally, rerooting trees with BS support values sometimes causes these values to end up with the wrong clade. This causes the wrong edges to be contracted, so solving this problem may also improve this filtering approach.

Biological interpretation

The phylogenetic networks constructed by the different algorithms for the various inputs were not entirely consistent in their topology, but they did give some insights into the evolutionary history of SARS-CoV-2. They all show a SARS-CoV-2 clade and a SARS clade and signs of recombination from these two clades into bat-SL-CoVZC45, with TriLoNet and TriL2Net indicating that this recombination originates specifically from a RaTG13/SARS-CoV-2 ancestor and HKU3-1. This is in line with research by Boni et al. (2020), stating that "progenitors of the RaTG13/SARS-CoV-2 lineage appear to have recombined with the Hong Kong clade (with inferred breakpoints at 11.9 and 20.8 kb) to form the CoVZXC21/CoVZC45-lineage", where the Hong Kong clade includes HKU3-1. Regarding the SARS-

CoV-2 clade, the MPL algorithm resulted in some aberrant topologies, but the other algorithms consistently show RaTG13 as the closest relative of SARS-CoV-2, in agreement with previous research (Zhou et al., 2020; Boni et al., 2020; Wang et al., 2020). Our results do not show any signs of recombination between SARS-CoV-2 and pangolin coronaviruses or of SARS-CoV-2 being a recombinant virus.

Future research

It should be noted that the resolution of the approach we used for the tree-based algorithms is limited by the size of the sequences used to construct the trees, so recombinations involving a small part of the genome will often not be detectable. One way to improve this would be to find possible recombination breakpoints and use them to divide the genome instead of the somewhat arbitrary blocks and genes. There exist programs that try to detect recombination breakpoints in an MSA, such as RDP (Martin et al., 2015) and CUTAL (Jones et al., 2019), but these programs cannot be used directly in our current workflow. RDP finds a very high number of possible breakpoints for this type of data and will therefore result in a large number of input trees and CUTAL is not suitable for sequences as large as coronavirus genomes. Nevertheless, if such a method could be incorporated in the workflow, it may also result in higher quality trees, which is crucial for improving the phylogenetic networks constructed by tree-based algorithms.

Lastly, the networks constructed by TriLoNet and TriL2Net indicate that the level-1 restriction results in different reticulations for different taxon selections. These algorithms might therefore not be as suitable for data from coronaviruses, because their frequent recombination could require high level networks. Future research might focus on constructing higher level networks, for example by implementing an algorithm to construct level-2 trinets from MSAs to use as input for TriL2Net. Another option would be to implement an already described algorithm which can combine multiple level-1 networks into level-2 networks (Murakami et al., 2019).

Acknowledgements

I thank Guido Grimm for sharing his data set and answering my questions about the data and Teun Boekhout for providing feedback on my Literature Review and Approach. I also thank Sander Borst and Norbert Zeh for solving bugs in the algorithms. Lastly, many thanks to my supervisors Leo van Iersel, Leen Stougie, Mark Jones and Steven Kelk for all their ideas, help, feedback and support during the project.

Availability

All data, results, code and Supplemental Information (Spreadsheets and Tables) from this project are publicly available at a Github repository: https://github.com/rosanneandrea/corona-networks The Tables from the Supplemental Information are included after the References as well.

References

[1] Kristian G. Andersen, Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes, and Robert F. Garry. "The proximal origin of SARS-CoV-2". In: *Nature Medicine* 26.4 (Apr. 2020), pp. 450–452. DOI: 10.1038/s41591-020-0820-9.

[2] Dennis A. Benson et al. "GenBank". In: *Nucleic Acids Research* 41 (D1 Nov. 26, 2012), pp. D36–D42. DOI: 10.1093/nar/gks1195.

[3] Maciej F. Boni et al. "Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic". In: *Nature Microbiology* (July 28, 2020). DOI: 10.1038/s41564-020-0771-4.

[4] Sander Borst, Leo van Iersel, Mark Jones, and Steven Kelk. "New FPT algorithms for finding the temporal hybridization number for sets of phylogenetic trees". In: *arXiv:2007.13615 [cs, q-bio]* (July 27, 2020). arXiv: 2007.13615.

[5] Jie Cui, Fang Li, and Zheng-Li Shi. "Origin and evolution of pathogenic coronaviruses". In: *Nature Reviews Microbiology* 17.3 (Mar. 2019), pp. 181–192. DOI: 10.1038/s41579-018-0118-9.

[6] Patrick Dawson, Mamunur Rahman Malik, Faruque Parvez, and Stephen S. Morse. "What Have We Learned About Middle East Respiratory Syndrome Coronavirus Emergence in Humans? A Systematic Literature Review". In: *Vector Borne and Zoonotic Diseases (Larchmont, N.Y.)* 19.3 (2019), pp. 174–192. DOI: 10.1089/vbz.2017.2191.

[7] Christian Drosten et al. "Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome". In: *New England Journal of Medicine* 348.20 (May 15, 2003), pp. 1967–1976. DOI: 10.1056/NEJMoa030747.

[8] Bradley Efron, Elizabeth Halloran, and Susan Holmes. "Bootstrap confidence levels for phylogenetic trees". In: *Proceedings of the National Academy of Sciences* 93.23 (Nov. 12, 1996), pp. 13429–13429. DOI: 10.1073/pnas.93.23.13429.

[9] Joseph Felsenstein. "Confidence limits on phylogenies: an approach using the bootstrap". In: *Evolution* 39.4 (July 1985), pp. 783–791. DOI: 10.1111/j.1558-5646.1985.tb00420.x.

[10] Joseph Felsenstein. *Newick Standard format*. URL: https://evolution.genetics.washington.edu/phylip/newicktree.html (visited on 10/01/2020).

[11] Joseph Felsenstein. *PHYLIP format*. URL: https://evolution.genetics.washington.edu/phylip/doc/sequence.html (visited on 10/01/2020).

[12] Rachel L. Graham and Ralph S. Baric. "Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission". In: *Journal of Virology* 84.7 (Apr. 1, 2010), pp. 3134–3146. DOI: 10.1128/JVI.01394-09.

[13] Guido Grimm and David Morrison. "Harvest and phylogenetic network analysis of SARS virus genomes (CoV-1 and CoV-2)". In: (Mar. 30, 2020). type: dataset. DOI: 10.6084/m9.figshare.12046581.v3.

[14] Daniel H. Huson and Celine Scornavacca. "Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks". In: *Systematic Biology* 61.6 (Dec. 1, 2012), pp. 1061–1067. DOI: 10.1093/sysbio/sys062.

[15] Mark Jones et al. "Cutting an alignment with Ockham's razor". In: *arXiv:1910.11041 [cs, q-bio]* (Oct. 24, 2019). arXiv: 1910.11041.

[16] Sjors Kole. "Constructing level-2 phylogenetic networks from trinets". In: (2020).

[17] Michael M.C. Lai and David Cavanagh. "The Molecular Biology of Coronaviruses". In: *Advances in Virus Research*. Vol. 48. Elsevier, 1997, pp. 1–100. DOI: 10.1016/S0065-3527(08)60286-9.

[18] Tommy Tsan-Yuk Lam et al. "Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins". In: *Nature* (Mar. 26, 2020). DOI: 10.1038/s41586-020-2169-0.

[19] Xiaojun Li et al. "Emergence of SARS-CoV-2 through recombination and strong purifying selection". In: *Science Advances* (May 29, 2020), eabb9153. DOI: 10.1126/sciadv.abb9153.

[20] Angela D. Luis et al. "A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special?" In: *Proceedings of the Royal Society B: Biological Sciences* 280.1756 (Apr. 7, 2013), p. 20122753. DOI: 10.1098/rspb.2012.2753.

[21] Hayes K.H. Luk, Xin Li, Joshua Fung, Susanna K.P. Lau, and Patrick C.Y. Woo. "Molecular epidemiology, evolution and phylogeny of SARS coronavirus". In: *Infection, Genetics and Evolution* 71 (July 2019), pp. 21–30. DOI: 10.1016/j.meegid.2019.03.001.

[22] David R. Maddison, David L. Swofford, and Wayne P. Maddison. "Nexus: An Extensible File Format for Systematic Information". In: *Systematic Biology* 46.4 (Dec. 1, 1997). Ed. by David Cannatella, pp. 590–621. DOI: 10.1093/sysbio/46.4.590.

[23] Wayne P. Maddison. "Gene Trees in Species Trees". In: *Systematic Biology* 46.3 (Sept. 1, 1997), pp. 523–536. DOI: 10.1093/sysbio/46.3.523.

[24] Darren P. Martin, Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. "RDP4: Detection and analysis of recombination patterns in virus genomes". In: *Virus Evolution* 1.1 (Mar. 2015). DOI: 10.1093/ve/vev003.

[25] Yukihiro Murakami, Leo van Iersel, Remie Janssen, Mark Jones, and Vincent Moulton. "Reconstructing Tree-Child Networks from Reticulate-Edge-Deleted Subnetworks". In: *Bulletin of Mathematical Biology* 81.10 (Oct. 2019), pp. 3823–3863. DOI: 10.1007/s11538-019-00641-w.

[26] James Oldman, Taoyang Wu, Leo van Iersel, and Vincent Moulton. "TriLoNet: Piecing Together Small Networks to Reconstruct Reticulate Evolutionary Histories". In: *Molecular Biology and Evolution* 33.8 (Aug. 2016), pp. 2151–2162. DOI: 10.1093/molbev/msw068.

[27] Etienne Simon-Loriere and Edward C. Holmes. "Why do RNA viruses recombine?" In: *Nature Reviews Microbiology* 9.8 (Aug. 2011), pp. 617–626. DOI: 10.1038/nrmicro2614.

[28] Alexandros Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* 30.9 (May 1, 2014), pp. 1312–1313. DOI: 10.1093/bioinformatics/btu033.

[29] Alexandros Stamatakis, Paul Hoover, and Jacques Rougemont. "A rapid bootstrap algorithm for the RAxML Web servers". In: *Systematic Biology* 57.5 (Oct. 2008), pp. 758–771. DOI: 10.1080/10635150802429642.

[30] Cuong Than, Derek Ruths, and Luay Nakhleh. "PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships". In: *BMC Bioinformatics* 9.1 (Dec. 2008), p. 322. DOI: 10.1186/1471-2105-9-322.

[31] Leo van Iersel, Remie Janssen, Mark Jones, Yukihiro Murakami, and Norbert Zeh. "A Practical Fixed-Parameter Algorithm for Constructing Tree-Child Networks from Multiple Binary Trees". In: *arXiv:1907.08474 [cs, math, q-bio]* (July 19, 2019). arXiv: 1907.08474.

[32] Leo van Iersel and Steven Kelk. "Constructing the Simplest Possible Phylogenetic Network from Triplets". In: *Algorithmica* 60.2 (June 2011), pp. 207–235. DOI: 10.1007/s00453-009-9333-0.

[33] Leo van Iersel and Vincent Moulton. "Trinets encode tree-child and level-2 phylogenetic networks". In: *Journal of Mathematical Biology* 68.7 (June 1, 2014), pp. 1707–1729. DOI: 10.1007/s00285-013-0683-5.

[34] Hongru Wang, Lenore Pipes, and Rasmus Nielsen. *Synonymous mutations and the molecular evolution of SARS-Cov-2 origins*. preprint. Evolutionary Biology, Apr. 21, 2020. DOI: 10.1101/2020.04.20.052019.

[35] Dingqiao Wen, Yun Yu, Jiafan Zhu, and Luay Nakhleh. "Inferring Phylogenetic Networks Using PhyloNet". In: *Systematic Biology* 67.4 (July 1, 2018). Ed. by David Posada, pp. 735–740. DOI: 10.1093/sysbio/syy015.

[36] Patrick C. Y. Woo, Yi Huang, Susanna K. P. Lau, and Kwok-Yung Yuen. "Coronavirus Genomics and Bioinformatics Analysis". In: *Viruses* 2.8 (Aug. 24, 2010), pp. 1804–1820. DOI: 10.3390/v2081803.

[37] Zi-Wei Ye, Shuofeng Yuan, Kit-San Yuen, Sin-Yee Fung, Chi-Ping Chan, and Dong-Yan Jin. "Zoonotic origins of human coronaviruses". In: *International Journal of Biological Sciences* 16.10 (2020), pp. 1686–1697. DOI: 10.7150/ijbs.45472.

[38] Yun Yu and Luay Nakhleh. "A maximum pseudo-likelihood approach for phylogenetic networks". In: *BMC Genomics* 16 (S10 Dec. 2015), S10. DOI: 10.1186/1471-2164-16-S10-S10.

[39] Ali M. Zaki, Sander van Boheemen, Theo M. Bestebroer, Albert D.M.E. Osterhaus, and Ron A.M. Fouchier. "Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia". In: *New England Journal of Medicine* 367.19 (Nov. 8, 2012), pp. 1814–1820. DOI: 10.1056/NEJMoa1211721.

[40] X. W. Zhang, Y. L. Yap, and A. Danchin. "Testing the hypothesis of a recombinant origin of the SARS-associated coronavirus". In: *Archives of Virology* 150.1 (Jan. 2005), pp. 1–20. DOI: 10.1007/s00705-004-0413-9.

[41] Peng Zhou et al. "A pneumonia outbreak associated with a new coronavirus of probable bat origin". In: *Nature* 579.7798 (Mar. 2020), pp. 270–273. DOI: 10.1038/s41586-020-2012-7.

Supplemental Tables

Name	**Original length**	**Length in MSA**
NSP1	539	539
NSP2	1913	1913
NSP3	5834	5521
NSP4	1499	1499
NSP5	917	917
NSP6	869	869
NSP7	248	248
NSP8	593	593
NSP9	338	338
NSP10	416	416
NSP11	38	38
NSP12	2794	2794
NSP13	1802	1802
NSP14	1580	1580
NSP15	1037	1037
NSP16	893	893
S	3821	2341
ORF3a	827	830
E	227	230
M	668	668
ORF6	185	192
ORF7a	365	368
ORF7b	131	134
ORF8	365	21
N	1259	1268
ORF10	116	116

Table 1: Overview of the genes for which trees were built, including their name/symbol, original length in the SARS-CoV-2 genome and the length of the corresponding sequence in the multiple sequence alignment. Order corresponds to their order in the SARS-CoV-2 genome. Note that all non-structural proteins (NSP's) together form the ORF-1a and ORF-1ab proteins, but their sequences are used as separate 'genes' here.

		Edge contraction thresholds					
		None	None	None	l=0.01	l=0.01	l=0.01
Breakpoints	**Selection**	None	s=70	s=90	None	s=70	s=90
Blocks	A (n=12)	9	9	9	9	9	9
	A- (n=11)	8	7	7	7	6	7
	B (n=9)	9	7	7	8	7	7
	B- (n=8)	5	5	5	5	5	4
	C (n=7)	5	5	4	5	5	4
	C- (n=6)	1	1	1	1	1	1
Genes	A (n=12)	23	22	21	19	19	20
	A- (n=11)	21	17	17	16	15	12
	B (n=9)	20	22	19	18	19	18
	B- (n=8)	12	9	9	10	8	8
	C (n=7)	10	9	9	9	8	8
	C- (n=6)	1	1	1	1	1	1

Table 2: Unique trees for different breakpoint location sets, taxon selections (n = number of taxa) and thresholds for edge contraction based on branch length (l) and BS support values (s). "–" indicates that no solution was found within the runtime limit of 5 minutes.

		Edge contraction thresholds					
		None	None	None	l=0.01	l=0.01	l=0.01
Breakpoints	**Selection**	None	s=70	s=90	None	s=70	s=90
Blocks	A (n=12)	2 1 1 0 0	4 2 1 1 1	2 0 0 0 0	5 5 6 5 6	4 2 3 2 1	1 1 1 0 0
	A- (n=11)	0 0 0 0 1	1 0 0 0 1	1 1 0 0 1	1 1 0 5 4	1 1 1 1 0	0 1 0 0 0
	B (n=9)	8 8 8 8 7	9 9 9 5 7	3 2 2 2 2	4 4 3 4 5	2 2 1 2 1	9 9 2 9 9
	B- (n=8)	2 1 1 1 0	3 3 3 0 0	5 5 6 5 5	1 1 1 0 1	1 1 0 1 0	1 1 1 1 0
	C (n=7)	6 6 6 1 1	4 4 2 4 4	11 11 11 8 9	1 3 2 2 1	5 5 9 8 11	4 2 2 2 1
	C- (n=6)	1 2 0 1 1	2 2 1 0 1	1 1 0 1 1	2 2 2 2 2	2 1 1 0 0	2 2 2 2 1
Genes	A (n=12)	7 6 0 0 0	1 3 3 1 0	5 5 3 4 3	2 2 3 3 2	3 2 2 3 3	0 1 8 0 1
	A- (n=11)	1 1 2 2 1	5 5 5 4 1	2 1 1 1 0	0 2 3 3 3	6 6 6 6 8	0 0 0 1 1
	B (n=9)	9 9 9 9 2	2 2 2 2 2	1 2 6 1 3	3 3 3 3 3	7 8 3 3 3	5 4 4 4 4
	B- (n=8)	0 0 0 2 3	4 1 1 1 0	5 2 5 2 3	1 1 1 0 1	1 1 1 1 2	1 0 1 0 0
	C (n=7)	1 2 1 3 1	5 5 5 1 1	2 6 2 3 2	7 6 6 6 2	3 8 8 5 8	3 3 3 4 4
	C- (n=6)	1 1 1 0 2	2 2 2 0 1	3 2 1 3 0	5 8 7 7 7	1 1 1 1 0	6 1 1 0 1

Table 3: Reticulation numbers of each of the five networks constructed by the Maximum Pseudo-Likelihood algorithm for different breakpoint location sets, taxon selections (n = number of taxa) and thresholds for edge contraction based on branch length (l) and BS support values (s).