



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



Methods for making multiple alignment of genomic sequences for severe acute respiratory syndrome coronavirus 2

Yoshiyuki Suzuki*

Graduate School of Science, Nagoya City University, Nagoya-shi, Aichi-ken 467-8501, Japan



ARTICLE INFO

Keywords:

Multiple alignment
Pairwise alignment
Phylogenetic analysis
Reference
Severe acute respiratory syndrome coronavirus
2

ABSTRACT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in December 2019 and caused a pandemic. To monitor the global transmission pattern of SARS-CoV-2, it is required to constantly update the phylogenetic tree of genomic sequences with 29.9 kb, which may be time consuming. Phylogenetic analysis of SARS-CoV-2 may be accelerated by making a multiple alignment of nucleotide sequences using the CPA (combining pairwise alignments) method, in which a pairwise alignment is made for a reference and each of other sequences, and the pairwise alignments are combined into a multiple alignment. Here it is shown from the analysis of 3729 genomic sequences for SARS-CoV-2 and outgroup strains that the CPA method can produce a multiple alignment with an elevated or a reduced number of variable sites depending on the reference compared to the OMA (ordinary multiple alignment) method, which was considered to be the most reliable. In particular, the topology of the phylogenetic tree constructed from the multiple alignment made using the CPA method adopting the outgroup sequence as the reference was considerably different from that using the OMA method, suggesting that the outgroup sequence may not be suitable as the reference in the CPA method.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged as the etiological agent of pneumonia called coronavirus disease 2019 (COVID-19) in Wuhan, China, in December 2019 (Zhu et al., 2020). Since then, SARS-CoV-2 caused a pandemic, and as of July 24, 2020, 15,012,731 cases and 619,150 deaths have been confirmed worldwide (World Health Organization, 2020). SARS-CoV-2 is classified in the genus *Betacoronavirus* of the family *Coronaviridae* (Gorbalenya et al., 2020; Lu et al., 2020). The virion of SARS-CoV-2 is enveloped and 50–200 nm in diameter (Chen et al., 2020), and is considered to utilize angiotensin-converting enzyme 2 (ACE2) as the cellular receptor (Lan et al., 2020; Shang et al., 2020; Wrapp et al., 2020; Yan et al., 2020). The genome of SARS-CoV-2 is a non-segmented, linear, and single-stranded RNA of positive polarity with the length of 29.9 kb (Wu et al., 2020; Zhou et al., 2020).

Phylogenetic analysis of genomic sequences for the strains of SARS-CoV-2 isolated over the world has been conducted to clarify the global transmission pattern in the course of the pandemic (Global Initiative on Sharing All Influenza Data, 2020). However, hundreds of genomes are daily sequenced, and thus the phylogenetic tree for thousands of genomic sequences is required to be constantly updated to reflect the real time situation of the pandemic. For the construction of phylogenetic tree, it is necessary to make multiple alignment of genomic sequences, which may be time consuming when the number of sequences

is large. Therefore, the alignment process has often been modified to accelerate the phylogenetic analysis of SARS-CoV-2; e.g., genomic sequences were partitioned into subsets, which were aligned separately with a reference and combined into a multiple alignment (Hadfield et al., 2018).

As the number of genomic sequences increases in the phylogenetic analysis of SARS-CoV-2, it may become critical to devise efficient ways to make multiple alignment. For this purpose, the approach using a reference may be extended such that a pairwise alignment is made for a reference and each of other sequences, and the pairwise alignments obtained are combined into a multiple alignment (Suzuki and Gojobori, 2001). The aim of the present study was to examine the property of this method in the phylogenetic analysis of SARS-CoV-2.

Genomic sequences for 15,419 and 1211 strains of SARS-CoV-2 were retrieved from Global Initiative on Sharing All Influenza Data (GISAID) (Shu and McCauley, 2017) (supplementary Table S1) and Virus Pathogen Database and Analysis Resource (ViPR) (Pickett et al., 2012) (supplementary Table S2), respectively, on May 1, 2020. After excluding the sequences containing < 29,500 nucleotide sites and ambiguous nucleotides, 3728 sequences were retained for the following analysis. Additionally, the genomic sequence for the bat coronavirus strain isolated from *Rhinolophus affinis*, RaTG13, which is known to be the most closely related to SARS-CoV-2, was retrieved from

* Corresponding author at.: Graduate School of Science, Nagoya City University, 1 Yamanohata, Mizuho-ku, Nagoya-shi, Aichi-ken 467-8501, Japan.
E-mail address: yossuzuk@nsc.nagoya-cu.ac.jp.

<https://doi.org/10.1016/j.mgene.2020.100785>

Received 10 June 2020; Received in revised form 24 July 2020; Accepted 14 August 2020

Available online 19 August 2020

2214-5400/ © 2020 Elsevier B.V. All rights reserved.

International Nucleotide Sequence Database (INSD) (accession number: MN996532) (Zhou et al., 2020) and used as the outgroup sequence for the phylogenetic analysis of SARS-CoV-2. In the phylogenetic analysis of nucleotide sequences, it is customary to use only the sites that are shared by all sequences in multiple alignment, to avoid problems caused by the heterogeneity in evolutionary rates among sites (Nei and Kumar, 2000). Therefore, multiple alignment without gaps was made for 3729 nucleotide sequences of SARS-CoV-2 and outgroup strains by each method as follows.

In the OMA (ordinary multiple alignment) method, multiple alignment for 3729 nucleotide sequences was made by including all sequences in the input file for the computer program MAFFT (version 7.305b) (Katoh et al., 2002), in which a multiple alignment was made progressively from closely related sequences to distantly related sequences along with a guide tree. The sites containing gaps were eliminated from the result. Although the OMA method was time consuming, this method was supposed to be the most reliable.

In the CPA (combining pairwise alignments) method, one of 3729 nucleotide sequences was selected as a reference. Pairwise alignment was made for the reference and each of other sequences with MAFFT (version 7.305b) (Katoh et al., 2002). The pairwise alignments were investigated to identify the nucleotide positions in the reference that were aligned without gaps to all other sequences. Multiple alignment was generated by aligning the nucleotides at the identified positions in the reference with the nucleotides at the corresponding positions in other sequences. The genomic sequence for the prototype strain of SARS-CoV-2, WH-Human 1 coronavirus (WHCV), which represented the ingroup strain in the phylogenetic analysis of SARS-CoV-2, was selected as the reference (Wu et al., 2020). The genomic sequence for RaTG13, which represented the outgroup strain, was also selected as the reference (Zhou et al., 2020). The CPA methods adopting WHCV and RaTG13 as the reference were called the CPA_{WHCV} and CPA_{RaTG13} methods, respectively, in the present study. Computations for making multiple alignments by the OMA, CPA_{WHCV}, and CPA_{RaTG13} methods were conducted on Mac OS X (version 10.10.5; 3.1 GHz Intel Core i7; 16 GB 1867 MHz DDR3).

The numbers of conserved and variable sites were counted for the multiple alignments made for 3729 nucleotide sequences of SARS-CoV-2 and outgroup strains using the OMA, CPA_{WHCV}, and CPA_{RaTG13} methods. Since the genetic variation among SARS-CoV-2 strains was effective for resolving the phylogenetic relationships among SARS-CoV-2 strains, the numbers of conserved and variable sites were also counted for 3728 sequences of SARS-CoV-2 strains.

Each of the multiple alignments made by the OMA, CPA_{WHCV}, and CPA_{RaTG13} methods was used for constructing the phylogenetic tree by the p distance-based neighbor-joining (NJp) method (Saitou and Nei, 1987) with MEGA (version 7.0.26) (Kumar et al., 2016), which has been reported to perform better than other methods generally for constructing the phylogenetic tree (Nei and Kumar, 2000; Yoshida and Nei, 2016). In each phylogenetic tree, the numbers of branches with the length of 0 (0-branches) and > 0 (non-0-branches) were counted separately. In addition, since the topology of the phylogenetic tree is constituted by interior branches (Nei and Kumar, 2000), the numbers of 0-branches and non-0-branches were divided into those of interior branches (interior 0-branches and non-0-branches, respectively) and exterior branches (exterior 0-branches and non-0-branches, respectively). The total branch length as well as the interior and exterior branch lengths was also computed for each phylogenetic tree. Furthermore, the topologies of the phylogenetic trees constructed from the multiple alignments made using the CPA_{WHCV} and CPA_{RaTG13} methods were compared with that made using the OMA method by examining the compatibility of partitions supported by interior non-0-branches.

Using the OMA method, multiple alignment for 3729 nucleotide sequences of SARS-CoV-2 and outgroup strains was made in 28,396 s. The multiple alignment contained 28,160 sites, among which 25,059 sites were conserved and 3101 sites were variable (Table 1). With

Table 1

Statistics for the multiple alignments for 3729 genomic sequences of SARS-CoV-2 and outgroup strains made using the OMA, CPA_{WHCV}, and CPA_{RaTG13} methods.

Sites	Method		
	OMA	CPA _{WHCV}	CPA _{RaTG13}
Conserved (SARS-CoV-2 only)	25,059 (25,955)	25,054 (25,950)	25,055 (25,948)
Variable (SARS-CoV-2 only)	3101 (2205)	3094 (2198)	3103 (2210)
Total	28,160	28,148	28,158

regard to 3728 sequences of SARS-CoV-2 strains, the numbers of conserved and variable sites were 25,955 and 2205, respectively. Multiple alignments were also made using the CPA_{WHCV} and CPA_{RaTG13} methods in 1132 and 1196 s, respectively. Both the numbers of conserved and variable sites were reduced to 25,950 and 2198, respectively, in the CPA_{WHCV} method (Table 1). On the other hand, the number of conserved sites was reduced to 25,948, but the number of variable sites was elevated to 2210 in the CPA_{RaTG13} method (Table 1).

In the phylogenetic tree constructed from the multiple alignment made using the OMA method, the number of 0-branches was 5088, consisting of 2498 interior and 2590 exterior 0-branches, whereas the number of non-0-branches was 2367, consisting of 1228 interior and 1139 exterior non-0-branches (Table 2; Supplementary Fig. S1) (Felsenstein, 1986). The total branch length of the phylogenetic tree was 0.13701, which was divided into 0.03404 and 0.10297 for the interior and exterior branch lengths, respectively. In the CPA_{WHCV} method, the numbers of interior and exterior non-0-branches were decreased to 1225 and 1135, respectively (Table 2; Supplementary Fig. S2). The interior and exterior branch lengths were also decreased to 0.03402 and 0.10256, respectively. In contrast, in the CPA_{RaTG13} method, the numbers of interior and exterior non-0-branches were increased to 1276 and 1156, respectively (Table 2; Supplementary Fig. S3). The interior and exterior branch lengths were also increased to 0.03459 and 0.10370, respectively.

Among 1225 interior non-0-branches constituting the topology of the phylogenetic tree constructed using the CPA_{WHCV} method, only 23 (2%) were incompatible to the topology of the phylogenetic tree constructed using the OMA method (Fig. 1). In contrast, 189 (15%) of 1276 interior non-0-branches constituting the topology of the phylogenetic tree constructed using the CPA_{RaTG13} method were incompatible to the topology of the phylogenetic tree constructed using the OMA method (Fig. 1). Therefore, the incompatibility in the topology of the phylogenetic tree constructed using the CPA_{RaTG13} method to that using the OMA method was significantly greater than that of the topology of the phylogenetic tree constructed using the CPA_{WHCV} method to that using the OMA method ($P = 1.17 \times 10^{-28}$ by χ^2 test).

In the present study, the property of the CPA method for making multiple alignment of nucleotide sequences was examined in comparison to the OMA method through the analysis of 3729 genomic sequences for SARS-CoV-2 and outgroup strains. Compared to the multiple alignment made using the OMA method, which was supposed to be the most reliable, the multiple alignment made using the CPA_{WHCV} method contained fewer variable sites, resulting in a decrease in the number of interior non-0-branches in the phylogenetic tree. In contrast, the multiple alignment made using the CPA_{RaTG13} method contained more variable sites, resulting in an increase in the number of interior non-0-branches in the phylogenetic tree. However, many of the interior non-0-branches in the phylogenetic tree constructed using the CPA_{RaTG13} method were incompatible with those in the phylogenetic trees constructed using the OMA and CPA_{WHCV} methods, which were mostly compatible with each other.

Generally, the rate of mis-alignment is higher in aligning distantly

Table 2

Statistics for the phylogenetic trees for 3729 genomic sequences of SARS-CoV-2 and outgroup strains constructed from the multiple alignments made using the OMA, CPA_{WHCV}, and CPA_{RaTG13} methods.

Category	Method		
	OMA	CPA _{WHCV}	CPA _{RaTG13}
0-branch [interior, exterior]	5088 [2498, 2590]	5095 [2501, 2594]	5023 [2450, 2573]
Non-0-branch [interior, exterior]	2367 [1228, 1139]	2360 [1225, 1135]	2432 [1276, 1156]
Total branch length [interior, exterior]	0.13701 [0.03404, 0.10297]	0.13657 [0.03402, 0.10256]	0.13829 [0.03459, 0.10370]

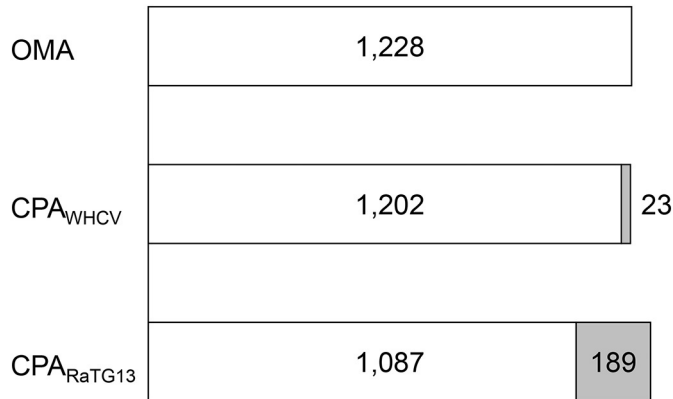


Fig. 1. Numbers of interior non-0-branches in the phylogenetic trees for 3729 genomic sequences of SARS-CoV-2 and outgroup strains constructed from the multiple alignments made using the OMA, CPA_{WHCV}, and CPA_{RaTG13} methods. The fractions of interior non-0-branches compatible and incompatible to the topology of the phylogenetic tree constructed using the OMA method were indicated with blank and gray bars, respectively.

related sequences than in aligning closely related sequences (Pollard et al., 2004). Thus, in the analysis of genomic sequences for 3729 SARS-CoV-2 and outgroup strains using the CPA_{WHCV} and CPA_{RaTG13} methods, the rate of mis-alignment was higher in the pairwise alignment of ingroup sequences than in the pairwise alignment of ingroup and outgroup sequences. In the CPA_{WHCV} method, in which an ingroup sequence was selected as the reference, 3727 pairwise alignments of ingroup sequences and one pairwise alignment of ingroup and outgroup sequences were conducted. On the other hand, in the CPA_{RaTG13} method, in which an outgroup sequence was selected as the reference, 3728 pairwise alignments of ingroup and outgroup sequences were conducted. Therefore, the rate of mis-alignment may be higher in the CPA_{RaTG13} method than in the CPA_{WHCV} method, as observed in the present study, suggesting that the outgroup sequence may not be suitable as the reference in the CPA method.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.mgene.2020.100785>.

Declaration of Competing Interest

The author declares no conflict of interest.

Acknowledgements

The author thanks two anonymous reviewers for valuable comments. This work was supported by JSPS KAKENHI Grant Number JP19K12221 and AMED Grant Number JP19fk0108033h0003 to Y.S.

References

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y.,

- Xia, J., Yu, T., Zhang, X., Zhang, L., 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *Lancet* 395, 507–513.
- Felsenstein, J., 1986. The Newick tree format. <http://evolution.genetics.washington.edu/phylip/newicktree.html>.
- Global Initiative on Sharing All Influenza Data, 2020. Next hCoV-19 App. <https://www.gisaid.org/epiflu-applications/next-hcov-19-app>.
- Gorbalenya, A.E., Baker, S.C., Baric, R.S., de Groot, R.J., Drosten, C., Gulyaeva, A.A., Haagmans, B.L., Lauber, C., Leontovich, A.M., Neuman, B.W., Penzar, D., Perlman, S., Poon, L.L.M., Samborskiy, D.V., Sidorov, I.A., Sola, I., Ziebuhr, J., 2020. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* 5, 536–544.
- Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., Neher, R.A., 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 4121–4123.
- Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T., 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066.
- Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874.
- Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., Wang, X., 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 581, 215–220.
- Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., Chen, J., Meng, Y., Wang, J., Lin, Y., Yuan, J., Xie, Z., Ma, J., Liu, W.J., Wang, D., Xu, W., Holmes, E.C., Gao, G.F., Wu, G., Chen, W., Shi, W., Tan, W., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* 395, 565–574.
- Nei, M., Kumar, S., 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford, New York.
- Pickett, B.E., Sadat, E.L., Zhang, Y., Noronha, J.M., Squires, R.B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C.N., Dietrich, J., Klem, E.B., Scheuermann, R.H., 2012. ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* 40, D593–D598.
- Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E., Eisen, M.B., 2004. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinform.* 5, 6.
- Saitou, N., Nei, M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.
- Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., Li, F., 2020. Structural basis of receptor recognition by SARS-CoV-2. *Nature* 581, 221–224.
- Shu, Y., McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro. Surveill.* 22, 30494.
- Suzuki, Y., Gojobori, T., 2001. Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b. *Gene* 276, 83–87.
- World Health Organization, 2020. Coronavirus Disease 2019. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>.
- Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S., McLellan, J.S., 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367, 1260–1263.
- Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E.C., Zhang, Y.-Z., 2020. A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269.
- Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., Zhou, Q., 2020. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* 367, 1444–1448.
- Yoshida, R., Nei, M., 2016. Efficiencies of the NJp, maximum likelihood, and Bayesian methods of phylogenetic construction for compositional and noncompositional genes. *Mol. Biol. Evol.* 33, 1618–1624.
- Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., Shi, Z.-L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273.
- Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W., 2020. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733.