MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220174

Phylogenetic Analyses of SARS-CoV-2 Strains Reveal Its Link to the Spread of COVID-19 Across the Globe

Shashata Sawmya[a,$], Arpita Saha[a,$], Sadia Tasnim[a,$], Naser Anjum[a], Md. Toufikuzzaman[a], Ali Haisam Muhamad Rafid[a], Mohammad Saifur Rahman[a], M. Sohel Rahman[a], Tanvir Alam[b]

[a] Department of CSE, Bangladesh University of Engineering and Technolog, West Palashi, Dhaka-1205, Bangladesh
[b] College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar
[$] Authors contributed equally

Abstract

This study leveraged the phylogenetic analysis of more than 10K strains of novel coronavirus (SARS-CoV-2) from 67 countries. Due to the requirement of high-end computational power for phylogenetic analysis, we leverage a fast yet highly accurate alignment-free method to develop the phylogenetic tree out of all the strains of novel coronavirus. K-Means clustering and PCA-based dimension reduction technique were used to identify a representative strain from each location. The resulting phylogenetic tree was able to highlight evolutionary relationships of SARS-CoV-2 genome and, subsequently, linked to the interpretation of facts and figures across the globe for the spread of COVID-19. Our analysis revealed that the geographical boundaries could not be explained by the phylogenetic analysis of novel coronavirus as it placed different countries from Asia, Europe and the USA in very close proximity in the tree. Instead, the commute of people from one country to another is the key to the spread of COVID-19. We believe our study will support the policymakers to contain the spread of COVID-19 globally.

Keywords:
COVID-19, Coronavirus, Phylogenetic tree

Introduction

COVID-19 is now considered as the biggest public health concern of this century [1, 2] and it has already surpassed the previous two outbreaks due to the coronavirus, namely, Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and Middle East Respiratory Syndrome Coronavirus (MERS-CoV). The virus acting behind this epidemic is known as Severe Acute Respiratory Syndrome Coronavirus 2 or in short SARS-CoV-2 virus. The novel coronavirus is a single stranded RNA virus having nearly 30K bases on average [3,4] and to understand the spread of COVID-19 disease, we need to investigate multiple strains of the novel coronavirus in as systematic fashion.

Traditionally alignment-based methods [5,6] are used to understand the phylogenetic relationship of the viral strains. However, these methods are computationally expensive and memory intensive. Consequently, alignment-free methods [7] are drawing attention in the scientific community to compare viral sequences as well as in constructing the phylogenetic relationship. As part of this study, an alignment-free phylogenetic analysis was carried out to uncover the evolutionary relationship among the strains of SARS-CoV-2 from multiple countries in the world. Based on our computational workflow, we identified a single representative strain from each location and built the phylogenetic tree to discover their evolutionary relationship. In the sequel, the resulting phylogenetic tree was linked to the spread of COVID-19 across the globe with evidences.

Methods

Data Collection and Preprocessing

We collected 10,179 SARS-CoV-2 genome sequences up to the date 24 April, 2020 (cut-off date) from the GISAID initiative dataset [8]. These are high quality complete viral genome sequences submitted by the scientific institutes of individual countries. These viral genome sequences were used for identifying the representative viral sequences (see next section below) and subsequent phylogenetic analyses. We note that, subsequently, more data have been made available under GISAID, but due to computational constraints and since we are focused on unearthing the initial spread, we restricted ourselves to the cut-off date above.

Identifying Representative Viral Sequence

We have used an alignment-free genome sequence comparison method as proposed by Li *et al.* in [9] as briefly described below. At first the sequence set is divided into subsets of sequences based on the location. All sequences are converted into representative 18-dimensional vector suggested in [9].

Pairwise distance among vectors derived from the fast vector method [9] were computed using Euclidean distance. Due to the high dimensionality of the resulting distance matrix, we resort to Principal Component Analysis (PCA) technique [10] to reduce the dimension of the matrix. Subsequently, we used K-means clustering [11] to identify the corresponding cluster centers. For the K-means clustering algorithm, we have used the implementation of [12] and used the default parameters except for the number of clusters which were set to 1 for determining the cluster center for each of the subsets. For each location-based cluster, the representative sequence (i.e., the "centroid" of the cluster) is then identified and used in the subsequent step of the pipeline. Figure 1 highlights the computational pipeline adopted for this study.

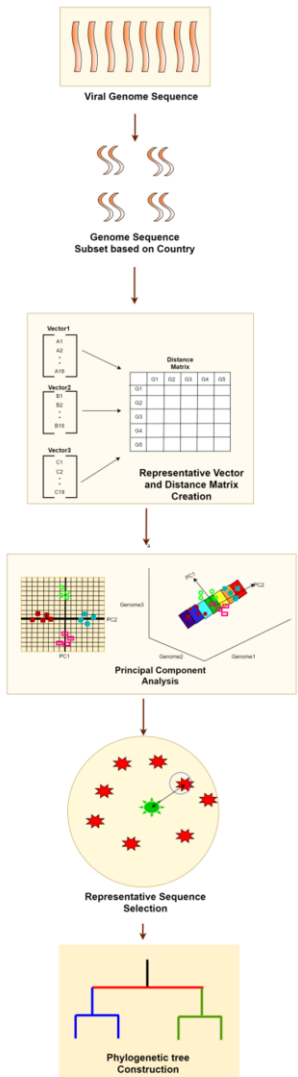

Figure 1 – Workflow of computational pipeline

Phylogeny Estimation

The evolutionary relationship among the representative sequences of different country-wise clusters has been estimated by constructing a phylogenetic tree. We have used the Neighbour Joining algorithm [13] for phylogenetic tree construction since it is more reliable [14]. We have used Euclidean distance among the vectors to prepare the distance matrix.

Results

As part of this study, we analyzed more than 10,000 SARS-Cov-2 strains from 67 countries. Table 1 highlights the top ten countries in terms of the number of strains. The details can be found as Supplementary File S1. From Table 1, we can observe that majority of the strains were collected from UK, USA, Australia.

Table 1– Strains per country from GISAID

Country	Strain	Country	Strain
UK	5414	Belgium	386
USA	2378	China	310
Australia	1045	France	226
Netherlands	567	Spain	146
Iceland	565	Canada	130

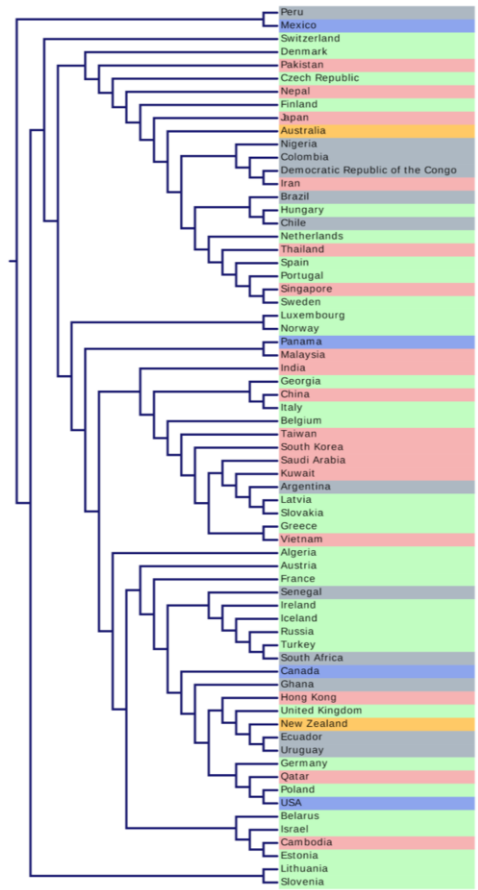

Figure 2 – The estimated phylogenetic tree of all the representative sequences. The representative sequence identified in the previous steps for each country were used in phylogeny estimation. In total 67 representative sequences from 67 countries were taken.

After applying K-means clustering and PCA techniques, we identified the representative sequence of each of the 67 countries as present in the GISAID dataset (upto cut-off date). The estimated phylogenetic tree was constructed from the representative sequences is shown in Figure 2. In what follows, we will be referring to this tree as the SARS-CoV-2 Phylogenetic Tree (SC2PT).

Discussion

Our analyses reveal a very close (evolutionary) relationship between the genome sequences of China and Italy (Figure 2). Similarity was found among the virus strains of the USA, Germany, Qatar and Poland. These countries have similar numbers of deaths and although not geographically directly adjacent (except for Germany and Poland) they have strong air connectivity among them.

In fact, a number of interesting relationships can be inferred from the SC2PT tree (Figure 2) as follows.

- The Italian strain of the virus is believed to be transmitted from China via two Chinese tourists [15]. This relationship is clearly portrayed in the SC2PT tree where the two strains appear to be immediate siblings.
- Poland's strain is in the same clade as that of Germany, which can be explained by the fact that its strain (through Poland's Patient Zero) came from Germany [16].
- Taiwan is geographically very close to China. The virus was confirmed to have spread to Taiwan on January 21, 2020, through a 50-year-old woman who had been teaching in Wuhan, China [17]. The virus strains from these regions are also close together as can be seen from the SC2PT tree, about 6 branches apart. Similar relationship can also be inferred from the tree between China and South Korea: the strain of the virus in South Korea is believed to be transmitted from China firstly through a 35-year-old Chinese woman and secondly by a 55-year old South Korean national [17]. Interestingly, from the SC2PT tree it can also be deduced that the South Korean strain is very close to that of Taiwan and also near to the strain from China. The incident of a Taiwanese woman being deported from South Korea after refusing to stay at a quarantine facility can be a probable explanation as to how the South Korean strain might have found its path to Taiwan [18].
- On March 2, 2020, the virus was confirmed to have reached Portugal, when it was reported that a Portuguese 33-year-old man working in Spain was tested positive for COVID-19 after returning home [19]. Subsequently, within a span of 9 days, 5 more cases were reported all originating from Spain [19][20]. The fact that the first cases of COVID-19 in Portugal originated from Spain is clearly captured in our SC2PT tree.
- The SC2PT tree suggests that India's strain is closely related to that from China and also Italy (around 4 branches) and that it is also connected to that from Saudi Arabia. These relationships can be explained as follows.
    a. On January 30, India reported its first case of COVID-19 in Kerala, which rose to three cases by February 3; all were students who had returned from Wuhan, China [21][22].
    b. On March 12, 2020, a 76-year-old man returning from Saudi Arabia became the first death case of the virus in the country [23]. India didn't impose a travel ban on Saudi Arabia at that point also.
    c. A Sikh preacher that returned from travel to Italy and Germany, carrying the virus, turned into a "super spreader" by attending a Sikh festival in Anandpur Sahib during 10–12 March [24][25].
- Strains from Austria and Greece are quite close to each other and near to that of Italy in the SC2PT tree as they are believed to be transmitted from Italy. The two people diagnosed with coronavirus in the Tyrol region of Austria were both Italian citizens [26] and a Greek woman who recently returned home from northern Italy became Greece's first coronavirus case [27][28].
- Turkey's first identified case was a man who was travelling in Europe [29]. Turkey also announced a huge number of cases and subsequent deaths, which were originating from Europe. In our inferred relationship, we can see that the Turkish representative strain is quite close to several Central and Western European countries like Russia, Iceland and Ireland which can be backed up by the two facts stated above.
- It is visible from the SC2PT tree that the strain of Germany is very close to the strains of both Poland and the USA. It might be the case that the community transmission occurred concurrently in both USA and Poland from Germany which hit the peak of pandemic before both USA and Poland [30].
- Qatar has the second highest number of COVID-19 patients in the Middle East [31]. The first case of Qatar was reported on February 27,2020 to be a man working in Iran [32]. Qatar introduced a travel ban to and from Germany and the USA as precautionary measures in Mid-March, quite a while later following the first occurrence. Qatar has 5 air-routes with Germany and USA, with more than 10 airlines operating in that route. Though the first case has originated from Iran, it might be the case that subsequent patients were found to be travelling from the aforementioned countries as a result of which the travel ban was introduced. Our estimated SC2PT tree places Qatar very close to both the USA and Germany.
- While we can certainly explain many of the relationships identified by the estimated SC2PT tree, there are some relationships which are not that apparent. One such example is the direct relationship between Vietnam and Greece. While apparently, there exists no direct relationship, when investigated further, we identified something interesting. Patient Zero of Greece is believed to have been contaminated during her trip to the Milan Fashion week which took place during February 18-24, 2020 [33]. Interestingly, the first COVID-19 patient in Hanoi [34] left Hanoi on February 15 to visit family members living in London, England and three days later, she traveled from London to Milan City. Could she be in contact with Patient Zero of Greece or any other who had been contaminated by the latter, before returning to London on February 20? We can't be certain, but our inferred relationship between Vietnam and Greece certainly put a lot of legitimacy to that question.

- Finally, we are unable to find any apparent explanation analyzing the reported news sources for a few other strong relationships inferred by the tree (e.g., Congo-Iran, Panama-Malaysia, Sweden-Singapore, Japan-Australia, etc.). This could be because of the inherent inaccuracies of the distance matrices as well as the limitations of the tree estimation algorithms: none of these algorithms are 100% accurate. From another angle, perhaps, the SC2PT did identify these relationships correctly; but the relevant incidences were not accurately identified or not documented.

Limitations

There are few limitations in this study. We analyzed nearly 10,200 SARS-CoV-2 strains that were avaialble in GISAID up to April, 2020. The strains were able to cover 67 countries in total. We only consider a single representative strain from each cluster, which may lead to the loss of information from other member of same cluster.

Conclusions

SC2PT tree is expected to reveal the evolutionary relationship of the viral strains. However, with careful scrutiny we have some apparently unusual but interesting observations. For example, we do not notice geographically adjacent countries in Europe as neighbours in SC2PT. In addition to that, although the USA and Canada share the longest un-militarized international border in the world, representative strains do not appear to be sister branches as they should have been. On the other hand, the relationships captured by SC2PT tree seems interesting on its own right. It is therefore hoped that this line of analysis would be useful for scientists to further investigate along this line and possible further genomic analysis could shed more light on this.

References

[1] Coronavirus Disease (COVID-19). *World Health Organization*, World Health Organization, www.who.int/emergencies/diseases/novel-coronavirus-2019.

[2] Gates, B. Responding to COVID-19 — A Once-in-a-Century Pandemic? *New England Journal of Medicine,* 382(18) 2020, 1677-1679.

[3] Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H. et. al (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *The Lancet*, 395(10224), 565-574.

[4] Wrapp D, Wang N, Corbett KS, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*. 2020;367(6483):1260-1263.

[5] Altschul, S. F. et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–402 (1997). [6]Larkin, M. A. *et al*. Clustal w and clustal x version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).

[7] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology* 2017;18(1):186.

[8] Elbe, S., and Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges,* 2017 1:33-46.

[9] Li, Y., He, L., Lucy He, R. *et al.* A novel fast vector method for genetic sequence comparison. *Sci Rep* **7,** 12226 (2017).

[10] Principal Component Analysis and Factor Analysis. (n.d.). *Principal Component Analysis Springer Series in Statistics,* 150-166.

[11] Hartigan, J. A., Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm.
Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1 (1979), pp. 100-108

[12] Sklearn.cluster.KMeans *Scikit*, scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[13] Saitou, N.; Nei, M. "The neighbor-joining method: a new method for reconstructing phylogenetic trees". *Molecular Biology and Evolution*. 1987, 4 (4): 406–425

[14] Mihaescu R, Levy D, Pachter L. "Why neighbor-joining works". *Algorithmica* 2009, 54 (1): 1–24.

[15] Online, Chiara Severgnini e Redazione. "Coronavirus, Primi Due Casi in Italia: Sono Due Turisti Cinesi." *Corriere Della Sera*, Corriere Della Sera, 31 Jan. 2020.

[16] Szymczak, Jakub. "Koronawirus w Lubuskiem. 44 Godziny, Dwa Razy Za Wolno. Daleko Do Laboratorium." Oko.press, Oko.press, 6 Mar. 2020, oko.press/koronawirus-wojewodztwo-lubuskie/.

[17] https://www.mk.co.kr/news/society/view/2020/01/80017/

[18] (LEAD) Taiwanese Woman Deported for Refusing to Stay at Quarantine Facility. *Yonhap News Agency*, 이민지, 6 Apr. 2020.

[19] Expresso. "Ministra Confirma Primeiro Caso Positivo De Coronavírus Em Portugal." *Jornal Expresso*, Expresso, 2 Mar. 2020, expresso.pt/sociedade/2020-03-02-Ministra-confirma-primeiro-caso-positivo-de-coronavirus-em-Portugal.

[20] "RELATÓRIO DE SITUAÇÃO". *Direção-Geral da Saúde* (in Portuguese). 2020-03-12.

[21] davyreid73. "India Confirms Its First Coronavirus Case." *CNBC*, CNBC, 30 Jan. 2020, www.cnbc.com/2020/01/30/india-confirms-first-case-of-the-coronavirus.html.

[22] https://weather.com/en-IN/india/news/news/2020-02-14-kerala-defeats-coronavirus-indias-three-COVID-19-patients-successfully

[23] India's First Coronavirus Death Is Confirmed in Karnataka, *Hindustan Times*, 12 Mar. 2020.

[24] Coronavirus: India 'Super Spreader' Quarantines 40,000 People, *BBC News*, BBC, 27 Mar. 2020, www.bbc.com/news/world-asia-india-52061915.

[25] Wallen, Joe. 40,000 Indians Quarantined after 'Super Spreader' Ignores Government Advice, *The Telegraph*, Telegraph Media Group, 28 Mar. 2020, www.telegraph.co.uk/news/2020/03/28/40000-quarantined-punjab-super-spreader-ignores-government-advice/.

[26] Renton, Adam, and Mike Hayes. Austria's 2 Coronavirus Cases Are Italian Citizens, *CNN*, Cable News Network, 26 Feb. 2020.

[27] Thomson Reuters, Greece Confirms First Coronavirus Case, a Woman Back from Milan. *Reuters*, 26 Feb. 2020

[28] Kambas, Michele. As Coronavirus Takes Hold, Greece Worries about Migrant Camps. *Reuters*, Thomson Reuters, 27 Feb. 2020.

[29] Turkey Remains Firm, Calm as First Coronavirus Case Confirmed. *Daily Sabah*, 10 Mar. 2020.
[30] Rourke, Alison. "Europe's Coronavirus Numbers Offer Hope as US Enters 'Peak of Terrible Pandemic'." *The Guardian*, Guardian News and Media, 6 Apr. 2020, www.theguardian.com/world/2020/apr/06/europes-coronavirus-numbers-offer-hope-as-us-enters-peak-of-terrible-pandemic .
[31] "Qatar to United States Flights." *Flights from Qatar*, www.qatar.to/United-States/Qatar-to-United-States.php.
[32] "Qatar Reports First Case of Coronavirus." *The Peninsula Qatar*, www.thepeninsulaqatar.com/article/29/02/2020/Qatar-reports-first-case-of-coronavirus.
[33] Natalie, et al. "Greece's 'Patient Zero' Shares Coronavirus Experience." *Greek City Times*, 1 May 2020, www.greekcitytimes.com/2020/05/01/greeces-patient-zero-shares-coronavirus-experience/.
[34] Person. Vietnam Confirms 17th COVID-19 Patient - VnExpress International. *VnExpress International – Latest News, Business, Travel and Analysis from Vietnam*, VnExpress.net, 10 Mar. 2020.

Supplementary File

Supplementary Files are shared in GitHub:
https://github.com/tanviralambd/COVID-PhylogenticTree

Note

An early version of this manuscript is available at BioRxiv: https://www.biorxiv.org/content/10.1101/2020.06.03.131987v1.full

Address for correspondence

Tanvir Alam, College of science and engineering, Hamad Bin Khalifa University, Doha, Qatar.
Email: talam@hbku.edu.qa; Phone: +974-4454-2277