Acta Mathematica Scientia, 2021, **41B**(3): 1017–1022
https://doi.org/10.1007/s10473-021-0323-x
©Innovation Academy for Precision Measurement Science and Technology, Chinese Academy of Sciences, 2021

Acta Mathematica Scientia
数学物理学报
http://actams.apm.ac.cn

ANALYSIS OF THE GENOMIC DISTANCE BETWEEN BAT CORONAVIRUS RATG13 AND SARS-COV-2 REVEALS MULTIPLE ORIGINS OF COVID-19*

Shaojun PEI (裴少君)
Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

Stephen S.-T. YAU (丘成栋)†
Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China
Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing 101408, China
E-mail: *yau@uic.edu*

Abstract The severe acute respiratory syndrome COVID-19 was discovered on December 31, 2019 in China. Subsequently, many COVID-19 cases were reported in many other countries. However, some positive COVID-19 samples had been reported earlier than those officially accepted by health authorities in other countries, such as France and Italy. Thus, it is of great importance to determine the place where SARS-CoV-2 was first transmitted to human. To this end, we analyze genomes of SARS-CoV-2 using k-mer natural vector method and compare the similarities of global SARS-CoV-2 genomes by a new natural metric. Because it is commonly accepted that SARS-CoV-2 is originated from bat coronavirus RaTG13, we only need to determine which SARS-CoV-2 genome sequence has the closest distance to bat coronavirus RaTG13 under our natural metric. From our analysis, SARS-CoV-2 most likely has already existed in other countries such as France, India, Netherland, England and United States before the outbreak at Wuhan, China.

Key words SARS-CoV-2; multiple origins of COVID-19; mathematical genomic distance; k-mer natural vector

2010 MR Subject Classification 92-08

1 Introduction

The severe acute respiratory syndrome COVID-19 was reported on December 31, 2019 in Wuhan (Hubei province, China) and is caused by a new type of coronavirus called SARS-CoV-2. SARS-CoV-2 is the seventh pathogenic coronavirus to human, and another six types of human coronaviruses are MERS-CoV, SARS-CoV, HCoV-229E, HCoV-HKU1, HCoV-NL63, and HCoV-OC43. Although SARS-CoV-2 has a lower mortality rate than SARS-CoV, it is

*Received February 25, 2021; revised March 10, 2021. This work was supported by Tsinghua University Spring Breeze Fund (2020Z99CFY044), Tsinghua University start-up fund, and Tsinghua University Education Foundation fund (042202008).

†Corresponding author

highly contagious and less detectable with a long incubation period [1]. So it is more threatening than other coronaviruses.

The early SARS-CoV-2's cases were associated with a sea food market in Wuhan. But its origin and intermediate host are still unclear. Bat coronavirus RaTG13 is the most similar sequence to SARS-CoV-2 found so far, which provides evidence for a bat origin of SARS-CoV-2 [2]. However, bat coronavirus RaTG13 was collected in 2013 and formed a distinct lineage from SARS-CoV-2, which could not transmit to humans directly. Then a series of studies on intermediate hosts have been conducted, including pangolin, mink and so on [3-5]. Another controversial issue is the place of the earliest human-to-human SARS-CoV-2 transmission. Although the earliest cases reported in other countries were generally in February 2020, more and more studies indicate that SARS-CoV-2 was spreading in December 2019 in France, Italy and the United States [6-8]. However, there are not complete sequences of these samples. So we hope to analyze the relationship of the existing sequences of SARS-CoV-2 to infer the early transmission of SARS-CoV-2 in human hosts.

Traditional methods to analyze the relationship of genome sequences are based on multiple sequence alignment (MSA). After alignment, a matrix of similarities between genome sequences will be given. But the similarity does not satisfy the triangular inequality property of mathematical distance [9]. So it cannot reflect the real biological distance of genome sequences. In this paper, we use a mathematical method called k-mer natural vector method to code the complete genome sequences with high quality in GISAID (https://www.gisaid.org/) as vectors in the Euclidean space [10, 11]. Then a new natural distance between the vectors is defined to measure the relationship of sequences. Based on the results, we conclude that before the outbreak at Wuhan, China, SARS-CoV-2 most likely has already existed in other countries such as France, India, Netherland, England and United States.

2 Materials and Methods

2.1 Dataset

All the complete genome sequences of SARS-CoV-2 were downloaded on GISAID until July 19, 2020. To ensure the accuracy of analysis, the low-quality sequences which contain letters other than A, C, G and T are eliminated from the dataset. Finally, there are 15,641 sequences in our dataset. The accession numbers of SARS-CoV-2s are shown in supplementary file 1. All the reference sequences of ss-RNA viruses were downloaded from NCBI up to March 23, 2020. In this study, we remove three types of sequences: (1) viruses without family label; (2) families including one or two sequences; and (3) viruses including letters other than A, C, G and T. Totally 2051 sequences are retained, which belong to 40 families. The details of these sequences are shown in supplementary 2.

2.2 K-mer Natural Vector

Definition 2.1 Let $S = s_1s_2s_3\dots s_n$ be a genomic sequence of length n, where $s_i \in \{A, C, G, T\}, i = 1, \dots, n$. K-mer is defined as a string of k consecutive nucleotides within a genomic sequence. For a given positive integer k, there are 4^k types of k-mers. Then the k-mer natural vector of the genomic sequence is composed of the following three components:

1. Let n_{l_i} denote the counts of k-mer l_i in S, where $i = 1, 2, \ldots, 4^k$.

2. Let $\mu_{l_i} = \frac{\sum_{j=1}^{n_{l_i}} v[l_i][j]}{n_{l_i}}$ specify the average location of k-mer l_i, where $v[l_i][m]$ is the distance from the first base to the j-th k-mer l_i in sequence S.

3. Let $D_m^{l_i} = \frac{\sum_{j=1}^{n_{l_i}} (v[l_i][j]-\mu_{l_i})^m}{n_{l_i}^{m-1} n^{m-1}}$ be the m-th central moment of position of k-mer l_i.

Then we can get the k-mer natural vector of the sequence S: $(n_{l_1}, \ldots, n_{l_{4^k}}, \mu_{l_1}, \ldots, \mu_{l_{4^k}}, D_2^{l_1}, \ldots, D_2^{l_{4^k}}, \ldots, D_m^{l_1}, \ldots, D_m^{l_{4^k}})$.

The correspondence between a genomic sequence and its associated k-mer natural vector is one-to-one and it is obvious that for any given k-mer l_i, higher central moments converge to zero quickly for a random generated sequence [11]. For example, for bat coronavirus RaTG13, the magnitude of $D_3^{l_i}$ is 10^{-3}, which is significantly smaller than that of $D_2^{l_i}$ of 10^3. Then the components of $D_3^{l_i}$ have little effect on the value of the Euclidean distance between k-mer natural vectors. Thus, we only calculate up to the second central moment in our experiment.

2.3 A new natural metric on the space of genome sequences

For a given k, each genomic sequence is associated with a 3×4^k-dimensional k-mer natural vector in the Euclidean space. In the previous study, most researches only consider one specific k to measure the distances between sequences [12,13]. Thus, one tricky problem is how to choose the value of k. However, we believe that the natural metric should involve all the k-mers for $k \geq 1$. So, we propose a new metric containing the information of all the k-mers for $k \geq 1$.

Definition 2.2 Let $d_k(v_1, v_2)$ be the Euclidean distance between two k-mer natural vectors v_1, v_2 of two genome sequences s_1, s_2 for $\forall k \geq 1$, then the new natural metric of two genome sequences s_1, s_2 is defined as $D_k(s_1, s_2) = d_1(v_1, v_2) + \frac{1}{2^2} d_2(v_1, v_2) + \frac{1}{3^2} d_3(v_1, v_2) + \ldots + \frac{1}{k^2} d_k(v_1, v_2)$.

Theorem 2.3 The new metric D_k satisfies three properties:

- Non-negativity: $D_k(s_1, s_2) \geq 0$
- Positivity: if $D_k(s_1, s_2) = 0$, then $s_1 = s_2$.
- Symmetry: $D_k(s_1, s_2) = D_k(s_2, s_1)$.
- Triangle inequality: $D_k(s_1, s_2) \leq D_k(s_1, s_3) + D_k(s_2, s_3)$.

Proof $\because \ \forall k \geq 1, \ d_k(v_1, v_2) \geq 0$.

$\therefore D_k(s_1, s_2) = d_1(v_1, v_2) + \frac{1}{2^2} d_2(v_1, v_2) + \frac{1}{3^2} d_3(v_1, v_2) + \ldots + \frac{1}{k^2} d_k(v_1, v_2) \geq 0$.

If $D_k(s_1, s_2) = 0$, then $d_i(v_1, v_2) = 0, \ i = 1, \ldots, k$, $\therefore v_1 = v_2$. According to the one-to-one correspondence between a genome sequence and its k-mer natural vector [11], then $s_1 = s_2$.

$$\begin{aligned} D_k(s_1, s_2) &= d_1(v_1, v_2) + \frac{1}{2^2} d_2(v_1, v_2) + \frac{1}{3^2} d_3(v_1, v_2) + \ldots + \frac{1}{k^2} d_k(v_1, v_2) \\ &= d_1(v_2, v_1) + \frac{1}{2^2} d_2(v_2, v_1) + \frac{1}{3^2} d_3(v_2, v_1) + \ldots + \frac{1}{k^2} d_k(v_2, v_1) \\ &= D_k(s_2, s_1). \\ D_k(s_1, s_3) &= d_1(v_1, v_3) + \frac{1}{2^2} d_2(v_1, v_3) + \frac{1}{3^2} d_3(v_1, v_3) + \ldots + \frac{1}{k^2} d_k(v_1, v_3) \\ &\leq d_1(v_1, v_2) + d_1(v_2, v_3) + \frac{1}{2^2} (d_2(v_1, v_2) + d_2(v_2, v_3)) \end{aligned}$$

$$
\begin{aligned}
&+\ldots+\frac{1}{k^2}(d_k(v_1,v_2)+d_k(v_2,v_3))\\
&=D_k(s_2,s_1)+D_k(s_2,s_3).
\end{aligned}
$$

□

The beauty of our new natural metric is that it contains information of the distributions from 1-mer to k-mer and is a mathematical metric for two genome sequences.

3 Results

3.1 The choice of the most accurate natural metric by the nearest neighborhood classification of ss-RNA virus

The definition of the new metric is $D_k = d_1 + \frac{1}{2^2}d_2 + \frac{1}{3^2}d_3 + \ldots + \frac{1}{k^2}d_k$, where d_k is the Euclidean distance between k-mer natural vectors of two genome sequences. But due to the limitation of capacity of computing, we cannot calculate too large value of k. So, all the reference sequences of ss-RNA viruses are used to determine which metric D_k we should choose. All the ss-RNA viruses belong to 40 families. For k from 1 to 11, we use new metric D_k as the distance to perform the nearest neighborhood classification of virus families. The results are illustrated in Figure 1 by black bars. We can see that the highest classification accuracy is 91.1%, when $k = 7$. For comparison, we also calculate the nearest neighborhood classification accuracies using d_1 to d_{11} respectively, which are shown by white bars in Figure 1. Obviously, our new natural metric is more accurate. So we choose D_7 in the next analysis of SARS-CoV-2 genome sequences.

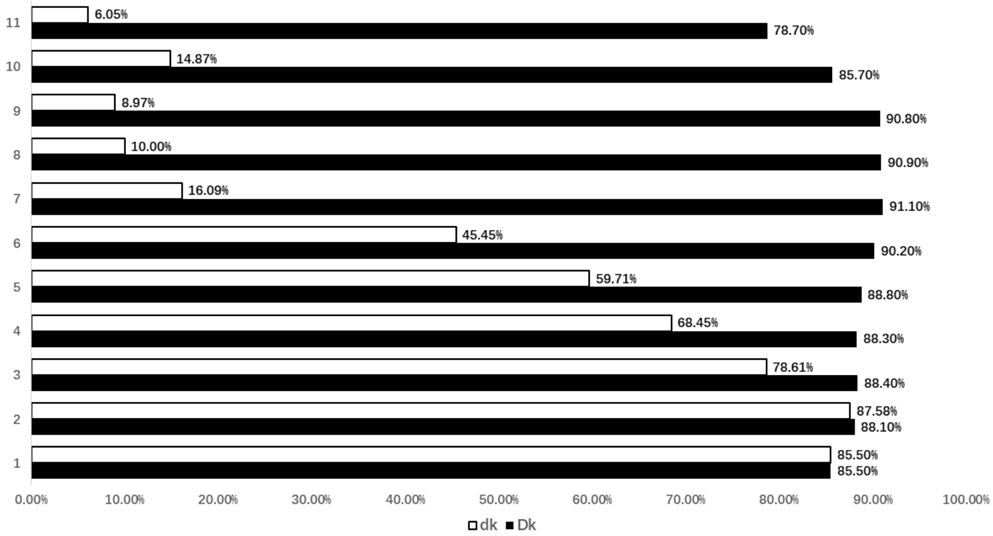

Figure 1 The classification accuracies of ss-RNA virus families for different k.

The accuracies by D_k are in black bars and the accuracies by d_k are in white bars.

3.2 The new natural metric between RaTG13 and SARS-CoV-2 genome sequences

In the previous study, bat coronavirus RaTG13 is the closest relative of SARS-CoV-2 [2]. So the distance between RaTG13 and each SARS-CoV-2 is calculated to analyze the transmission of SARS-CoV-2 in human hosts based on our new natural metric. According to the classification accuracies above, we choose $k = 7$ to calculate our new metric. The distances D_7 between the genome sequence of RaTG13 and all the genome sequences of SARS-CoV-2 in our dataset are ranked. The first five SARS-CoV-2 genome sequences with the shortest distance are shown in Table 1, which were collected in France, India, Netherlands, England and United States respectively. The distances of other sequences are shown in supplementary file 3. The distance between SARS-CoV-2 collected in Wuhan and bat coronavirus RaTG13 is 31006.95, ranking 426. This means that the SARS-CoV-2 genome sequences in Table 1 collected from these 5 countries are more similar with bat coronavirus RaTG13 than that of SARS-CoV-2 collected in Wuhan. These results indicate that the place where human-to-human SARS-CoV-2 transmission first happened is extremely unlikely to be Wuhan, but France, India, Netherlands, England and United States, with an accuracy rate higher than 91%.

Table 1 The top five genome sequences of SARS-CoV-2 with the shortest distance D_7

Rank	Accession Number	Distance D_7
1	hCoV-19/France/B5434/2020|EPI_ISL_443279|2020-04-01	30788.39
2	hCoV-19/India/S30/2020|EPI_ISL_455659|2020-05-01	30813.26
3	hCoV-19/Netherlands/Utrecht_10024/2020|EPI_ISL_454773|2020-03-26	30817.25
4	hCoV-19/England/BRIS-12E36C/2020|EPI_ISL_482027|2020-04-22	30818.11
5	hCoV-19/USA/CruiseA-1/2020|EPI_ISL_413606|2020-02-17	30818.19
426	hCoV-19/Wuhan/WH01/2019|EPI_ISL_406798|2019-12-26	31006.95

4 Discussion

Since December 2019, the severe respiratory pneumonia COVID-19 has spread globally. However, the first cases of COVID-19 could be earlier than those officially reported in many countries. Many studies have detected SARS-CoV-2 in earlier preserved biological or environmental samples. For example, Sridhar et al. suggested that SARS-CoV-2 may have appeared in the United States in December 2019 by identifying SARS-CoV-2-reactive antibodies. Of the 7,389 samples, 106 were reactive by pan Ig. And it failed to confirm whether these positive tests came from community transmission or travel transmission, because only 11 of the volunteers who donated blood have been to Asia recently [7]. Carrat et al. reported that SARS-CoV-2 infection may have occurred as early as November 2019 in France based on the anti-SARS-CoV-2 IgG test [8]. So the accurate identification of the origin of SARS-CoV-2 is a very important problem. However, the sequences in their studies are not complete. They cannot be analyzed by our method. In this paper, we not only provide a novel metric to study viral sequence based on k-mer natural vector, but also apply it to the analysis of the existing complete genome sequences of SARS-CoV-2 to identify the early circulation of SARS-CoV-2.

The previous methods based on k-mer always only consider the frequencies of k-mers and a certain value of k [12,13]. Here, the k-mer natural vectors contain both the frequency and the distribution of k-mers in the genome sequences. The correspondence between genome sequence and its k-mer natural vector is one-to-one. Especially, the k-mers for any k are involved in our new defined metric, so it can reflect the real biology distance between genome sequences and does not lose any information.

It is commonly accepted that SARS-CoV-2 is originated from bat coronavirus RaTG13, and the SARS-CoV-2 reference genome (NC_045512.2) [14] is uncertain whether it is earlier than the emerging strains. So we choose bat coronavirus RaTG13 as the reference and determine which SARS-CoV-2 genome sequence has the closest distance to bat coronavirus RaTG13 under our natural metric. According to the rank of distances, before the outbreak at Wuhan, SARS-CoV-2 most likely has already existed in other countries such as France, India, Netherlands, England and United States. So our result shows that Wuhan is extremely unlikely to be the first place of human-to-human SARS-CoV-2 transmission.

Acknowledgements We thank the researchers worldwide who sequenced and shared the complete genomes of SARS-CoV-2 and other coronaviruses from GISAID (https://www.gisaid.org/).

References

[1] Guan W, Ni Z, Yu H, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. New England Journal of Medicine, 2020, **382**: 1708–1720

[2] Zhou P, Yang X L, Wang X G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 2020, **579**: 270–273

[3] Lam T T Y, Jia N, Zhang Y W. et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature, 2020 **583**: 282–285

[4] Munnink B B O, Sikkema R S, Nieuwenhuijse D F, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Science, 2020, **371**(6525): eabe5901

[5] Dong R, Pei S, Yin C, et al. Analysis of the hosts and transmission paths of SARS-CoV-2 in the COVID-19 outbreak. Genes, 2020, **11**(6): 637

[6] Deslandes A, Berti V, Tandjaoui-Lambotte Y, et al. SARS-CoV-2 was already spreading in France in late December 2019. International Journal of Antimicrobial Agents, 2020, **55**: 106006

[7] Sridhar V B, Monica E P, Kacie G, et al. Serologic testing of U.S. blood donations to identify SARS-CoV-2-reactive antibodies: December 2019-January 2020. Clinical Infectious Diseases, 2020, ciaa1785

[8] Carrat F, Figoni J, Henny J, et al. Evidence of early circulation of SARS-CoV-2 in France: findings from the population-based "CONSTANCES" cohort. European Journal of Epidemiology, 2021. https://doi.org/10.1007/s10654-020-00716-2

[9] Yu C, He R L, Yau S S T. Protein sequence comparison based on K-string dictionary. Gene, 2013, **529**: 250–256

[10] Wen J, Chan R H F, Yau S -C, et al. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene, 2014, **546**: 25–34

[11] Deng M, Yu C, Liang Q, et al. A Novel Method of Characterizing Genetic Sequences: Genome Space with Biological Distance and Applications. PLoS ONE, 2011, **6**(3): e17293

[12] Sims G E, Jun S R, Wu G A, et al. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proceedings of the National Academy of Sciences, 2009, **106**: 2677–2682

[13] Sims G E, Jun S R, Wu G A, et al. Whole-genome phylogeny of mammals: evolutionary information in genic and non-genic regions. Proceedings of the National Academy of Sciences, 2009, **106**: 17077–17082

[14] Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature, 2020, **579**(7798): 265–269