Similarity Comparisons of SARS-CoV-2 Samples between Wuhan and G20

Zhenhui Wang, Mu Qiao, Jeffrey Zheng

Abstract To analyze SARS CoV-2 genomes, this paper uses a viral genome collected in Wuhan as a basic sequence to select a set of genomes from G20 countries/regions for comparison. Two methods are applied in visualization followed in the A1 and C1 modules of the MAS to provide 1D similarity projection and information entropy approaches. For a pair of two genomes segmented as M segments to calculate base differences corresponding to each segment, a measuring vector can be generated to be projected as a visual map. Refined information will be provided to calculate the information entropy corresponding to base differences. This provides quantitative measures to organize these genomes arranged into a linear order. The first method uses a line chart representation to illustrate relevant differences among genomes 1 by 1 between Wuhan and other regions. The second method uses a histogram representation to observe the information entropy projection results. Significant similarities are observed among the differences with approximately equal entropies with higher similarity. Based on the comparison of gene sequence diagrams between Wuhan and various regions, a difference analysis was carried out.
Keywords: metagenomic analysis system MAS, visual comparison, difference analysis, information entropy, visualization

Zhenhui Wang
School of Software Yunnan University e-mail: 1875154386@qq.com

Mu Qiao
Yunnan University e-mail: qiaomu@mail.ynu.edu.cn

Jeffrey Zheng
Key Laboratory of Quantum Information of Yunnan
Key Laboratory of Software Engineering of Yunnan
School of Software, Yunnan University e-mail: conjugatelogic@yahoo.com

This work was supported by the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018ZI002).

Background of The Research

SARS-CoV-2 is the seventh coronavirus known to infect humans [1,2]. SARS-CoV-2 is now widely spread throughout the world. As of April 26, 2020, more than 2.8 million cases have been infected worldwide. SARS CoV-2 first detected in Wuhan, which made some countries mistakenly believe that the disease originated in China, but the disease first appeared in China, which does not mean that it originated in China [3]. From the perspective of scientific research, tracing the source of SARS-CoV-2, conducting relevant data research and analysis, and having the exact research results, we can make a conclusion about the source of SARS-CoV-2.

Aim of The Research

Based on the viral genome of Wuhan, this paper analyzes the differences between the viral genome sample data of Wuhan and G20 countries/regions, finds countries or regions with high similarity to the Wuhan viral genome, and makes corresponding diagrams for reference by other researchers.

Description of The Data

For the convenience of observation and analysis, this paper divides the G20 regions and other regions into three parts, namely, A, B, and C parts. Part A includes A1: United States, Belgium, Brazil, South Korea, A2: France, Germany, United Kingdom, Singapore, A3: India, Australia; Part B is Japan; Part C includes C1: Turkey, Mexico, Italy, C2: China, Canada, C3: South Africa, Saudi Arabia, Russia, Chile. Note: Because the length of the Indonesian gene sequence is short and cannot be compared, this article uses Singapore data; to compare the data in China with Wuhan, this article uses Guangzhou data [4]. All genomes are collected from the GISAID GenBank: https://www.gisaid.org

Method of The Research

The paper uses two methods. The first method is the difference analysis method [5,6], which uses the A1 method 1D similarity projection principle of the architecture of the metagenomic analysis system MAS. The difference analysis method is to segment the two gene sequences with the same segment length, count the number of bases A, T, C, and G for each segment, and then make a difference in the number of bases in the corresponding segment. After taking the absolute value of the difference, the difference of bases A, T, C, and G is accumulated, and finally,

the total difference of the two gene sequences in each segmented base is obtained, and a comparison chart is drawn according to the calculation [7,8]. The difference formula is as follows.

$$S = \sum_{i=1}^{n} (|A_i - a_i| + |T_i - t_i| + |C_i - c_i| + |G_i - g_i|) \tag{1}$$

The second method is based on the information entropy analysis method [9]-[12], which uses the C1 method of the MAS. The specific method is the segmentation of two gene sequences, and the ratio of the difference between a certain base on the corresponding segment and the total difference is taken as the probability of the information entropy. Calculate the entropy on each segment, and then add the summation. The result is the information entropy of the difference between the two gene sequences. The information entropy of each sequence is arranged in accordance with the size of the projection on the coordinate system for easy observation. According to the information entropy value, select some country series for further comparison. The information entropy is formulated as follows.

$$H(U) = E[-\log p_i] = -\sum_{i=1}^{n} p_i \log p_i \tag{2}$$

Results and Analysis

According to the two methods, the graphical results and corresponding analysis are drawn as follows. Note: The abscissa of the graph is the number of segments M, and the ordinate is the difference.

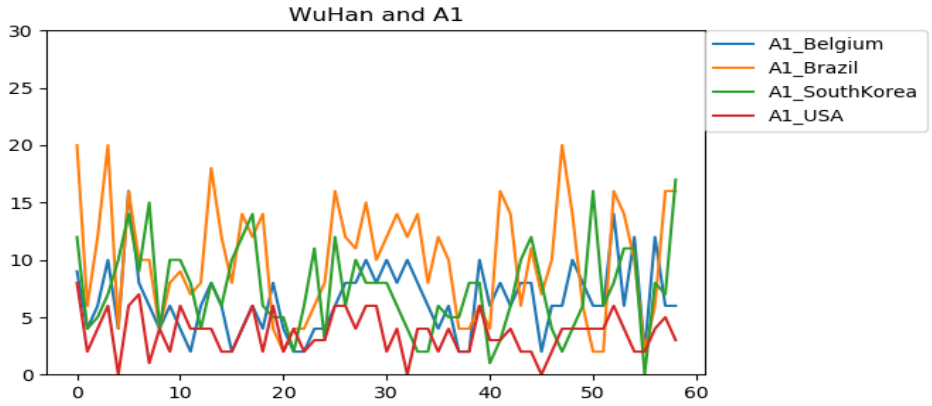

Fig. 1 Wuhan compared with Belgium, Brazil, South Korea and USA in A1

In Fig. 1, the virus sequence differences between Wuhan and the United States are smaller and similar in each segment, while the differences between Wuhan and Brazil are larger.

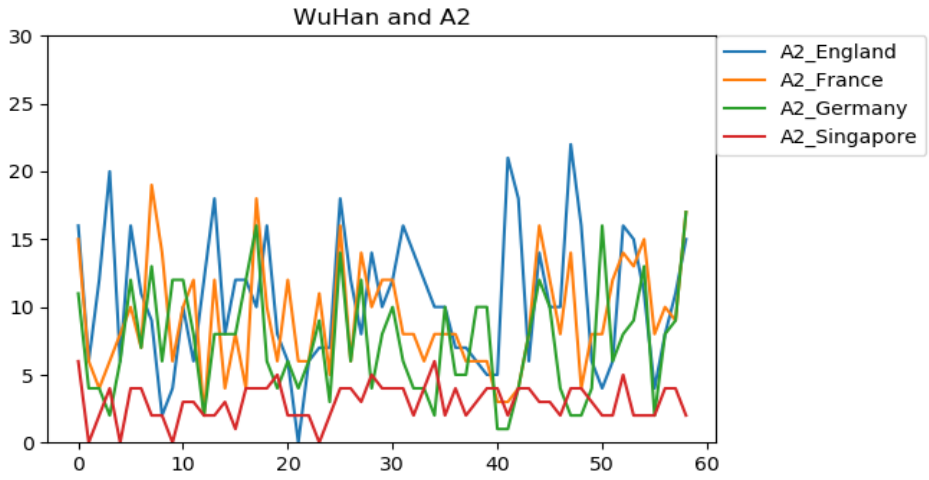

Fig. 2 Wuhan compared with England, France, Germany and Singapore in A2

In Fig. 2, the virus sequence differences between Wuhan and Singapore are very small and similar in each segment. In addition, the virus sequences in Wuhan and the United Kingdom have large fluctuations, with small differences in some segments and large differences in others.

The following figure is a graph drawn using the information entropy method. Ten countries in Part A of the G20 area are selected. The information entropy has been sorted from small to large in the figure. Its abscissa is the country, and the countries represented by the serial number are 1: Singapore, 2: United States, 3: India, 4: Australia, 5: Belgium, 6: France, 7: Germany, 8: United Kingdom, 9: South Korea, and 10: Brazil. The ordinate is information entropy.

Use the A1 method to make further sample similarity comparisons on the results of information entropy, as shown in the figure below.

Conclusion

The visual comparison of SARS CoV-2 genomes in Wuhan and G20 shows that the difference between virus gene sequences in Wuhan and some countries and the overall information entropy value are smaller, and the gene sequences have greater similarity, such as the United States and Singapore. It can also be concluded that the viral gene sequences of some countries are similar. For example, Australia and

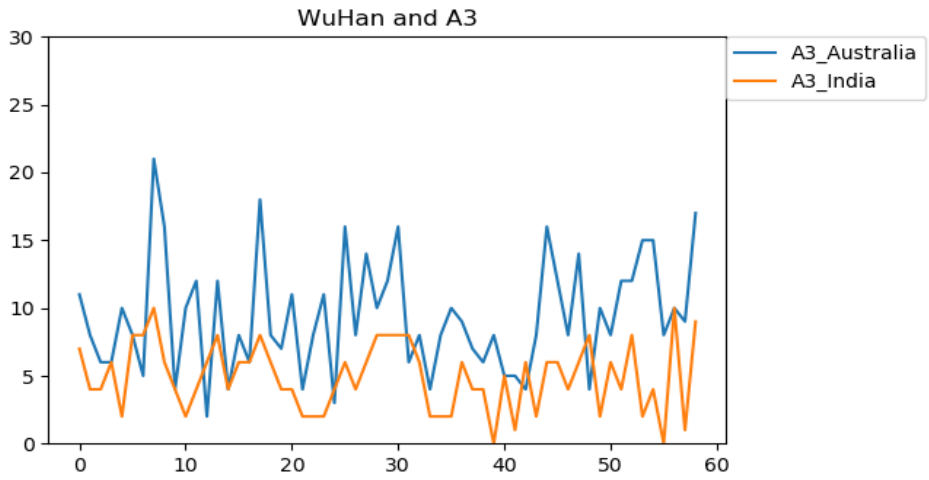

Fig. 3 The figure shows that the viral gene sequences of Wuhan and India have small differences in certain segments, with partial similarities, and large differences with Australia.

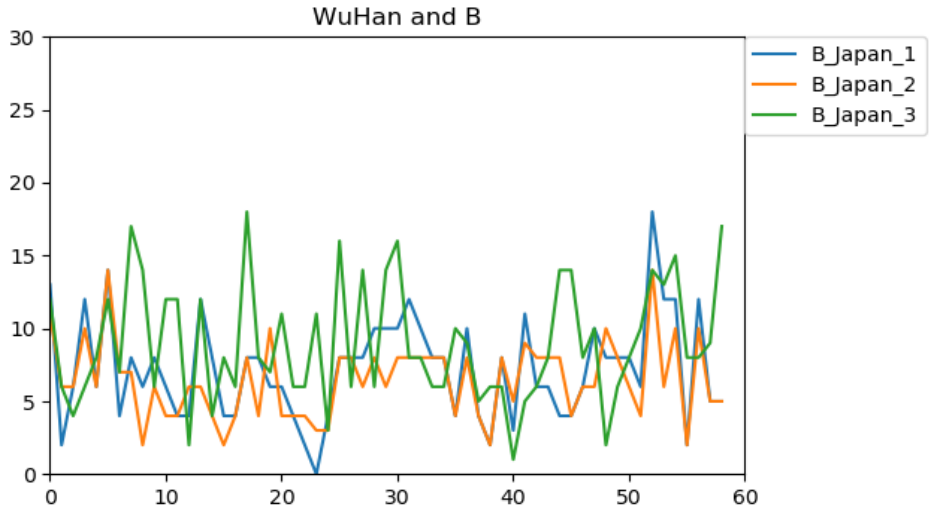

Fig. 4 Part B uses three Japanese viral gene sequences to compare with Wuhan. From the figure, it can be seen that the values of the three curves are not the same. The difference between Wuhan and B_Japan_1,B_Japan_2 is small, and most of them are similar, while the difference between Wuhan and B_Japan_3 is larger. It shows that the virus gene sequences of Wuhan and B_Japan_1 and B_Japan_2 are relatively similar.

France have similar sequences. From the graph of information entropy, can obtain the overall information entropy in Wuhan and part A of G20 regions and analyze the overall similarity between Wuhan and these countries. It can be seen which countries

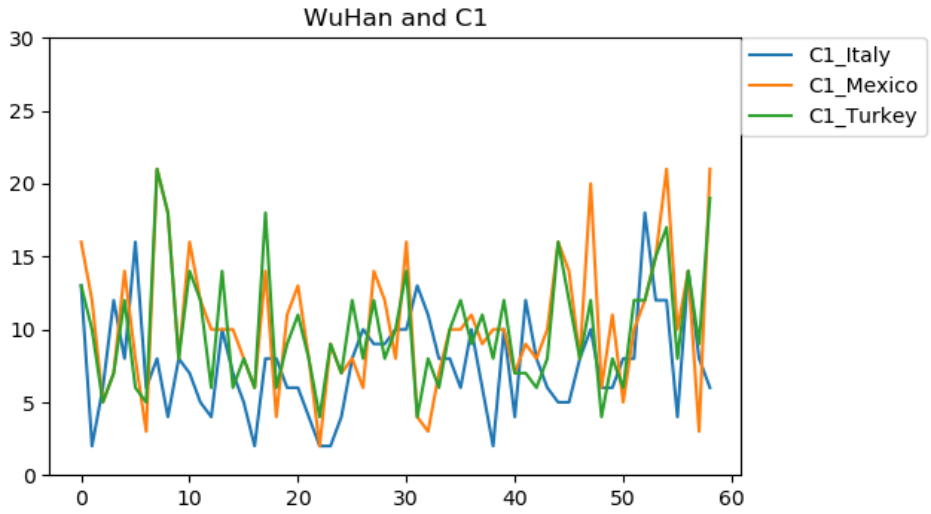

Fig. 5 Compared with the viral gene sequences of Wuhan and C1, most of the segments have large differences, and only the individual segments have small differences.

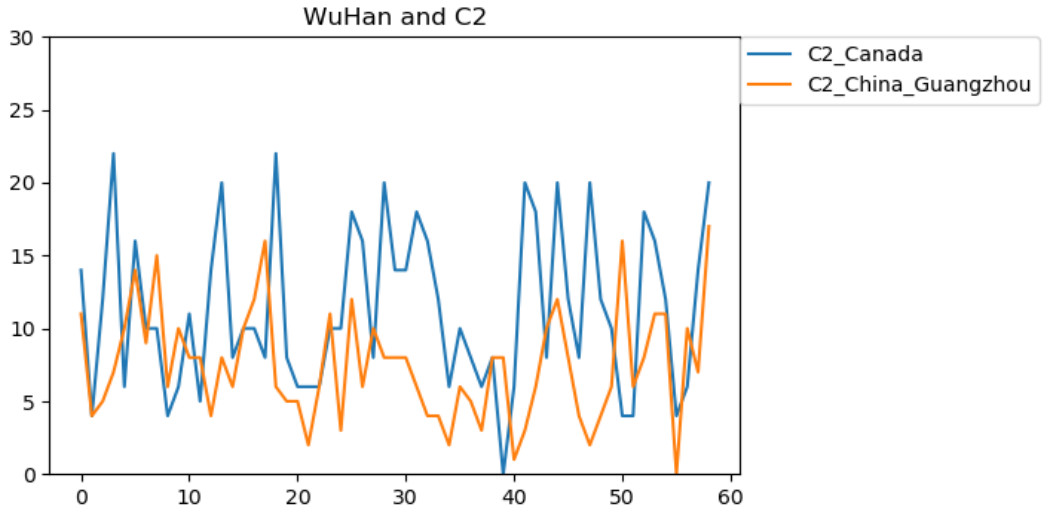

Fig. 6 The figure shows that the viral gene sequences of Wuhan and Canada differ greatly in most segments, and the overall difference is large. In addition, the difference between the virus gene sequences in Wuhan and Guangzhou is large, and the similarity is very small, indicating that the viruses in the two regions are not homologous.

have similar information entropy values to compare the sequence similarity of these countries.

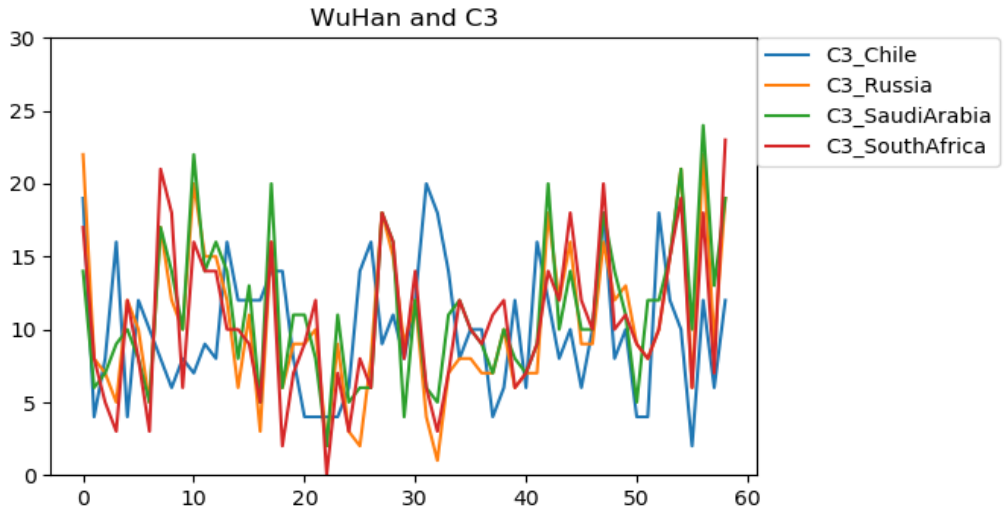

Fig. 7 The difference between most of the virus gene sequences in the Wuhan and C3 areas is large, only the difference between the middle sections is small, and the overall difference is large. However, the three difference curves of Wuhan and Russia, Saudi Arabia and South Africa partially overlap, indicating that the viral gene sequences of these three countries are relatively similar.

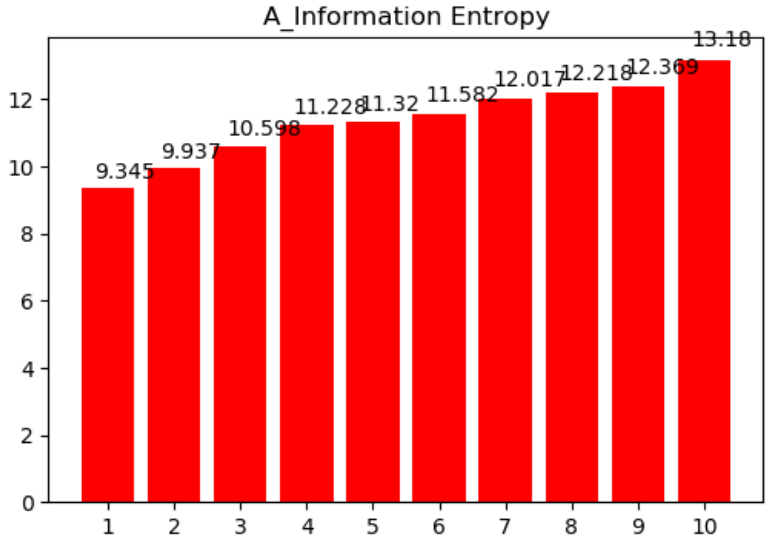

Fig. 8 From the figure, it is obvious that the virus gene sequence similarity between Part A and Wuhan is Singapore, followed by the United States and India and Brazil with the lowest similarity. Using the method of information entropy can more conveniently observe the overall result.

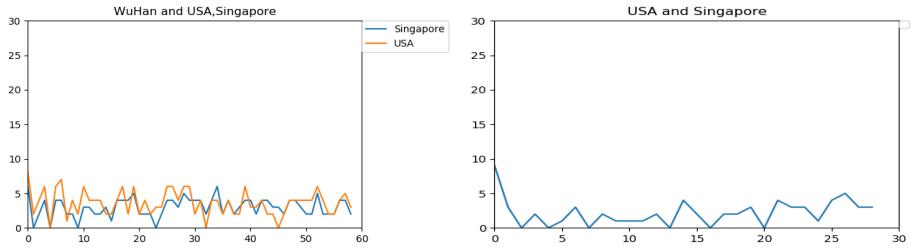

Fig. 9 The graph on the left compares Wuhan and Singapore, United States, using the A1 method. The graph on the right compares the difference between Singapore and the United States. It is concluded that the difference between these two countries is small and the similarity is high.

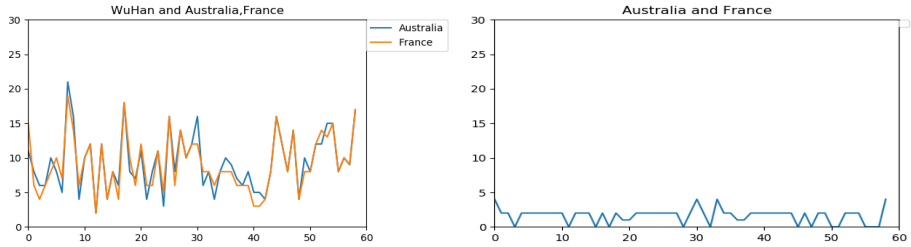

Fig. 10 The picture on the left is a comparison between Australia and France and Wuhan. It is found that the difference curve between the two countries and Wuhan is very close, and the information entropy values of the two countries are also very similar, so compare Australia and France. The figure on the right shows that the difference between the two countries is very small, indicating that the viral gene sequences of the two countries are very similar.

Conflict Interest

No confilct of interest has claimed.

Acknowledgements The authors would like to thank NCBI, GISAID, CNGBdb, Nextstrain for providing invaluable information on the newest dataset collections of SARS CoV-2 & other coronavirus genomes to support this project working smoothly.

References

1. Andersen, K.G., Rambaut, A., Lipkin, W.I. et al,*The proximal origin of SARS-CoV-2*, Nat Med 26, 450452 (2020).
2. Dilcher Meik,Werno Anja,Jennings Lance C, *SARS-CoV-2: a novel deadly virus in a globalised world*, The New Zealand medical journal,2020,133(1510).
3. Antoine Flahault,*Has China faced only a herald wave of SARS-CoV-2?*, The Lancet,2020(prepublish)

4. Zhanwei Du, Lin Wang, Simon Cauchemez, Xiaoke Xu, Xianwen Wang, Benjamin J. Cowling, and Lauren Ancel Meyers, *Risk for Transportation of 2019 Novel Coronavirus Disease from Wuhan to Other Cities in China*, Emerging Infectious Diseases,2020,26(5).
5. Michael Yarus, Life From An RNA World, Harvard University Press 2010
6. R. Durrentt, Probability Models for DNA Sequence Evolution, Springer 2008
7. Jeffrey Zheng, *Variant Construction from Theoretical Foundation to Applications*, Springer Nature 2019 https://www.springer.com/in/book/9789811322815
8. Jeffrey Zheng, *Variant Construction Theory and Applications, Vol.1: Theoretical Foundation and Applications*,Science Press 2020 (Chinese, Formal Publishing Soon).
9. Ryan J P, *Information, entropy and various systems*, Journal of theoretical biology,1972,36(1).
10. Day Troy, *Information entropy as a measure of genetic diversity and evolvability in colonization*, Molecular ecology,2015,24(9).
11. Dehghanzadeh Houshang,Ghaderi-Zefrehei Mostafa, Mirhoseini Seyed Ziaeddin,Esmaeilkhaniyan Saeid, Haruna Ishaku Lemu, Amirpour Najafabadi Hamed, *A new DNA sequence entropy-based Kullback-Leibler algorithm for gene clustering*, Journal of applied genetics,2020,61(2).
12. Yang Fan,Wu Duzhi,Lin Limei,Yang Jian,Yang Tinghong,Zhao Jing, *The integration of weighted gene association networks based on information entropy*, PloS one,2017,12(12).