Visualizations of Topologic Entropy on SARS-CoV-2 Genomes in Multiple Regions

Mu Qiao

Yunnan University

Renyang Liu

Yunnan University

Zhenhui Wang

Yunnan University

Xinmei Li

Yunnan University

Jeffrey Zheng (✉ conjugatelogic@yahoo.com)

Yunnan University <https://orcid.org/0000-0003-4225-7077>

Research Article

Keywords: sequence comparison, metagenomic analysis system MAS, SARS-CoV-2, viral genomics, topological entropy, diagram

Posted Date: January 11th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-65305/v3>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Visualizations of Topologic Entropy on SARS-CoV-2 Genomes in Multiple Regions

Mu Qiao, Renyang Liu, Zhenhui Wang, Xinmei Li, Jeffrey Zheng

Abstract The outbreak of novel coronavirus (SARS-CoV-2) developed into a global pandemic in a few months. The latest study found that the virus belongs to the beta coronavirus family. SARS-CoV-2 is highly similar to Pangolin CoV and BatCoV RaTG. Advanced scientific studies help traceability and vaccine development. In addition to the subgenus classification analysis of the virus, it is interesting for further exploration to focus attention on mutations and their transmissions in different regions. New mutations may be likely to affect the symptoms of the disease and the effectiveness of vaccination. This paper is focused on the study to make error bars and scatter graphs with the support of the metagenetic analysis system MAS. Using SARS-CoV-2 genomes in different countries and regions as input datasets, topological entropy values provide global characteristic quantities based on C_4 module for visualization. Sample results show that the method is powerful and useful for consistently integrating all genomes on one unique genomic index map. Various countries have confirmed their specific positions and projections under topologic entropies. Further explorations are required.

Keywords: sequence comparison, metagenomic analysis system MAS, SARS-CoV-2, viral genomics, topological entropy, diagram

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

Jeffrey Zheng
Yunnan University, Kunming e-mail: conjugatelogic@yahoo.com

This work was supported by the NSFC (62041213), the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018ZI002)

Introduction

Novel severe acute respiratory syndrome coronavirus (2019-nCoV, SARS-CoV-2) [1], has developed into a global pandemic [2]. Seven coronaviruses are known to cause human infection, of which four HCoV229E, HCoVNL63, HCoVHKU1, and HCoVOC43 usually cause cold symptoms in immune individuals. Other SARS-CoV (Severe Acute Respiratory Syndrome Coronavirus) and MERS-CoV (Middle East Respiratory Syndrome Coronavirus) originate from zoonosis and cause severe respiratory diseases and deaths [3]. It has been found that the new coronavirus SARS-CoV-2 belongs to the beta coronavirus family [4], which includes the previously mentioned severe acute respiratory syndrome (SARS) and the Middle East respiratory syndrome (MERS). Guo et al. proposed a virus-host prediction method based on deep learning [5] to detect viruses with a DNA sequence as input to predict their potential infection hosts. SARS-CoV-2 is more closely related to other human coronaviruses, especially SARS-CoV, Bat-SARS and MERS-CoV.

Through genomic analysis, SARS-CoV-2 comes from nature, because its genome is highly similar to bat coronavirus, and bats may be its natural host [6]. The ten 2019-nCoV genomic sequences obtained by Lu and Zhao from 9 patients are very similar, showing more than 99.98% sequence identity. It is worth noting that 2019-nCoV is closely related to two bat-derived (SARS)-like coronaviruses bat-SLCoVZC45 and bat-SL-CoVZXC21 collected in Zhoushan, eastern China in 2018 (identity is 88%), and the distance of the same virus family is SARS-CoV (approximately 79%) and MERS-CoV (approximately 50%) [7]. Because SARS-CoV-2 belongs to the same family as SARS-CoV and MERS-CoV, they have many similarities. The structure and pathogenicity of SARS-CoV-2 are very similar to that of SARS-CoV. Compared with SARS, SARS-CoV-2 spreads faster than SARS, but its lethality is relatively lower [8]. Therefore, Flynn protease inhibitors may be a potential drug therapy for SARS-CoV [9]. Other studies have found those novel coronavirus is closely related to severe acute respiratory syndrome-like coronavirus (bat-SL-CoVZC45 and bat-SL-CoVZXC21) derived from two species of bats [10].

Some studies suggest that the Pangolin-CoV, carried by pangolin, is likely to provide a natural gene pool for novel coronavirus. At the genome-wide level, the similarity between Pangolin-CoV and SARS-CoV-2 and BatCoV RaTG is 91.02% and 90.55%, respectively. The relationship between the S1 protein of pangolin-CoV and SARS-CoV-2 is much closer than that between the S1 protein and RaTG13 [11]. The results of a recent study [12] have shown that the critical atomic interaction between the spinous process protein receptor-binding domain (RBD) of SARS-CoV-2 and the host receptor angiotensin-converting enzyme 2 (ACE2) usually helps to regulate the spread of COVID-2019 across species and humans.

Aim of the study

As of April 26, 2020, the number of epidemic countries has reached 209, with more than 1.8 million confirmed cases and more than 200 thousand deaths. The epidemic has not been adequately controlled worldwide, and the number of confirmed cases is still increasing [13]. The study of virus sequence characteristics is significant for further virus classification, surveillance and vaccine development and treatment.

At present, most studies focus on the classification and analysis of the subgenus of the virus and find the closest virus cluster, which provides the direction for virus traceability and vaccine development and treatment. However, there is a lack of individual population analysis of the SARS-CoV-2 virus in different regions. Because RNA depends on RNA polymerase (RdRp) and has a high mutation rate [14], new mutations are likely to affect the symptoms of the disease and the efficacy of vaccination. Li found 19 lethal new mutations in novel coronavirus. She pointed out that novel coronavirus has modifications that can affect pathogenicity. Drug and vaccine research and development also need to consider these mutation factors [15]. It has been observed that novel coronaviruses are sensitive to the environment by mutating and producing different variants. SARS-CoV-2 will gradually adapt to the local climate, and mutations in the virus genome play a vital role in the process of transmission. The analysis of genome characteristics shows that there is a strong relationship between sample collection time, sample location and genetic diversity accumulation. The study found 116 mutations, some of which affected the severity and transmission of SARS-CoV-2.

In this study, the C_4 output module of the MAS are implemented, and the topological entropy is used as the feature of the sequence. We have made diagrams of the distribution of novel coronavirus RNA sequences in multiple regions [18]. The complete series of several SARS-Cov-2 genomes in the different areas were selected to study the relationship between geographical distribution and sequence characteristics of virus genomes. Finally, the developmental clusters of the strains are identified [19] so that the scientific and diagnostic communities associated with the coronavirus can benefit from it.

Material and Methods

Data

All the sequence data used in this study were downloaded from complete genomic sequences with NCBI and GISAID, with data size requirements of approximately 30000bp. The quality of the genome is high, to reduce the existence of unknown nucleotides as much as possible to avoid affecting the accuracy of the experiment. The total number of all sequences is 1337. The statistics of different countries are

shown in Table 1. The statistics of various cities in different regions involved in the investigation are shown in Table 2, Table 3 and Table 4.

Table 1 Statistical table of virus sequences in different countries

Country	Number	Country	Number	Country	Number
Belgium	218	China	200	Brazil	14
Italy	17	USA	675	Australia	60
France	26	Japan	92	Canada	35

Table 2 Statistical table of virus sequences in different regions of the United States

State/city	Number	State/city	Number	State/city	Number
Alaska	6	Arizona	5	California	11
Connecticut	8	Florida	9	Georgia	12
Illinois	6	Iowa	7	New Hampshire	4
Wisconsin	69	Massachusetts	19	Minnesota	10
New York	188	Pennsylvania	6	South Carolina	7
Utah	15	Virginia	16	Washington	210
North Carolina	10				

Table 3 Statistical table of virus sequences in different regions of China

State/city	Number	State/city	Number	State/city	Number
Beijing	10	Chongqing	2	Foshan	3
Guangdong	8	Guangzhou	6	Hangzhou	37
Jiangsu	4	Nanchang	10	Shandong	3
Shanghai	42	Shangrao	8	Shenzhen	5
Wuhan	29				

Topological Entropy

Information entropy theory is an essential tool in bioinformatics, and it is widely used in genome sequence analysis. Topological entropy (TE) is a kind of informa-

Table 4 Statistical tables of virus sequences in different regions of Australia, Italy and France

Australia		Italy		France	
State/city	Number	State/city	Number	State/city	Number
Victoria	18	Abruzzo	2	Grand-Est	2
Queensland	16	Friuli Venezia Giulia	4	Hauts de France	9
Western Australia	9	Rome	6	Ile de France	15
Sydney	17	Veneto	2		

tion entropy. Kirillova [16] uses the exponential growth rate that produces topological entropy to describe topological entropy.

Topological entropy can distinguish simulated gene sequences from natural gene sequences. Koslick [17] proposed using the approximate calculation of topological entropy to analyze gene sequences. He selects the fragment length of n and w by analyzing the maximum point of $P_w(n)$, to obtain an exponential growth rate, which is closest to the complexity of the sequence. Defined as

$$\lim_{n \rightarrow \infty} \frac{\log_2(P_{w^{4n+n-1}}(n))}{n} \quad (1)$$

In the formula, w represents the sequence length, and n is the sequence substring length, and the value of n in $4 \leq n \leq 16$ is obvious. $P_{w^{4n+n-1}}(n)$ represents the number of types of n words appearing in the sequence, which is also the k -mers commonly used in molecular biology. In this study, the sequence length w we used is approximately 30000bp, and the calculated n value is 7 and the n value is 8 for the experiment.

Topological Entropy Generation

TE is the C_4 part of four modules $\{CE, IE, ME, TE\}$. The process of TE generation is to input N sequences [20], calculate the average length of each sequence, select the best K -mers length, calculate the non-weight quantity of K -mers as the intermediate quantity [21], and finally obtain the topological entropy of the sequence [22]. The specific process is as follows.

1. According to the virus sequence length of $|w|$, the word length n can be determined.
2. Therefore, we get all the n -mers of this sequence, the total number is $|w| - n + 1$.
3. Filter out the repeated n -mers and get the number of n -mers, $P_w(n)$.

4. The topological entropy $H_W(n)$ of the sequence is calculated by substituting the formula.

5. Finally, the topological entropy of a country is calculated, and the average and standard deviation of the topological entropy of all sequences under that country are taken to make an error bar graph. And according to the topological entropy value of $n = 7$ and $n = 8$ of each sequence, as the characteristic quantity of the sequence, respectively correspond to the horizontal and vertical coordinates of the scatter diagram. Make a scatter diagram.

Results and Discussions

Make data error bar diagrams and scatter diagrams of different countries, and select the data with city (state) information to make diagrams of different regions.

Fig. 1(a)-(f) provide error bar diagrams of topological entropy for different countries. The topological entropy of the sequence of $n = 7$ is selected as the ordinate.

From the comparison between Fig 1(a) and the actual data, it is not difficult to find that the average topological entropy of different countries is approximately $H(w) = 0.9560$, and the data topological entropy of 0.9560 is likely to be an initial feature of SARS-CoV-2. In Fig. 1(b), we can observe that the entropy of the genome in most parts of the United States is higher than 0.9560, corresponding to Fig. 1(a); only a few areas are smaller, such as Washington, Minnesota, Wisconsin. The highest entropy value of New York is $H(w) = 0.9561751051408733$, and the lowest entropy value of $H(w) = 0.9558286454522853$ is the area with the most significant error change in the statistical field. There is an exciting graphic phenomenon in Fig. 1(c). In Wuhan, where coronavirus was first discovered in China, the entropy value of the mathematical sequence is the largest and the error range is also the largest. Fig. 1(d) and Fig. 1(e) Australia and Italy also show similar characteristics to Wuhan, and the error range of the maximum average entropy is also the largest. In Fig. 1(b), the United States shows the most significant margin of error in Illinois. However, in Fig. 1(d), the average entropy value of Italian city Abruzzo is as high as 0.9640, which is the maximum entropy value at present, and the entropy error range of this city is also the largest. Fig. 1(f) shows that the entropy of the virus sequence found in three French cities is 0.95599, the entropy value between different cities is stable, and the entropy error is $1/10000$.

Fig. 2(a)-(f) shows the scatter distribution of topological entropy in different countries or regions. Both $n = 7$ and $n = 8$ are selected as entropy calculation parameters, and the horizontal and vertical coordinates are entropy attributes.

In Fig. 2(a), some scattered points overlap and classify in different countries, such as the United States, Belgium, Australia, and China, their statistical aspects overlap and some mathematical points in the United States are separated. Next, specific to different cities in each country for observation. Comparing Fig. 2(b) and (a), the separation of statistical points in the United States is related to the sequence of New York. In Fig. 2(b), the mathematical entropy points of New York are more character-

(a)

(b)

(c)

(d)

(e)

(f)

Fig. 1 Error bar diagram of topological entropy in six maps (a)-(f); (a)different countries; (b) USA; (c) China; (d) Australia;(e)Italy; (f)France

Fig. 2 scatter distribution of topological entropy in six maps (a)-(f); (a)different countries; (b) USA; (c) China; (d) Australia;(e)Italy; (f)France

istic than other locations. In Fig. 2(c), the entropy points of Wuhan are widely distributed, which mostly coincide with those of other cities, while the entropy points of Hangzhou are separated. Fig. 2(d) shows the entropy scatter distribution in Australia. The sequence entropy values of Queensland and Victoria gather together, and the sequence entropy values of Sydney and Western Australia overlapped. Fig. 2(e) and (f) are more evenly distributed.

Conclusion

From the entropy value diagram of different countries and different regions, the method of using topological entropy to deal with the virus sequence is effective. First, the average entropy value of each country is stable, and second, it is observed that there are differences in the entropy value of different cities. The entropy richness and consistency of the viruses at the first discovery site is better. The aggregation and classification of entropy values in different regions were observed through the entropy scatter diagram.

Conflict Interest

No conflict of interest has been claimed.

Acknowledgements The authors would like to thank NCBI, GISAID, and 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. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, *The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2*, Nat Microbiol, 2020;5(4):536544. doi:10.1038/s41564-020-0695-z .
2. Haski ū z M, Kili S, Sara F, *Coronaviruses and SARS-COV-2*, Turk J Med Sci, 2020;50(SI-1):549556. Published 2020 Apr 21. doi:10.3906/sag-2004-127.
3. Khailany RA, Safdar M, Ozaslan M, *Genomic characterization of a novel SARS-CoV-2*, published online ahead of print, 2020 Apr 16, doi:10.1016. genrep.2020.100682
4. Zhou P, Yang XL, Wang XG, et al, *A pneumonia outbreak associated with a new coronavirus of probable bat origin*, Nature, 2020;579(7798):270273. doi:10.1038/s41586-020-2012-7.
5. Qian Guo, Mo Li, Chunhui Wang, Peihong Wang, Zhencheng Fang, Jie tan, Shufang Wu, Yonghong Xiao, Huaiqiu Zhu, *Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm*, doi: <https://doi.org/10.1101/2020.01.21.914044>.
6. Guo YR, Cao QD, Hong ZS, et al, *The origin transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak - an update on the status*, Mil Med Res, 2020;7(1):11. Published 2020 Mar 13. doi:10.1186/s40779-020-00240-0.

7. R Lu, X Zhao, J Li, et al, *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding*, Lancet (2020). published online Jan 30.
8. Chan JF, Yuan S, Kok KH, et al, *familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster*, Lancet. 2020;395(10223):514523. doi:10.1016/S0140-6736(20)30154-9.
9. Rabaan AA, Al-Ahmed SH, Haque S, et al, *SARS-CoV-2, SARS-CoV, and MERS-COV: A comparative overview*, Infez Med, 2020;28(2):174184.
10. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR, *Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges*, Int J Antimicrob Agents, 2020;55(3):105924. doi:10.1016/j.ijantimicag.2020.105924.
11. Zhang T, Wu Q, Zhang Z, Tarjan and D. Woods, *Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak*, Curr Biol, 2020;30(7):13461351.e2. doi:10.1016/j.cub.2020.03.022.
12. Wan Y, Shang J, Graham R, Baric RS, Li F, *Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus*, J Virol, 2020;94(7):e00127-20. Published 2020 Mar 17.
13. Dong E , Du H , Gardner L. Stanley, *An interactive web-based dashboard to track COVID-19 in real time*, Lancet Infect Dis 2020 Feb 19. PubMed PMID: 320871 14. Epub 2020/02/23
14. Duffy S, Shackelton LA, Holmes EC, *Rates of evolutionary change in viruses: Patterns and determinants*, Nature Reviews Genetics, 2008; 9: 267276. pmid:18319742.
15. Hangping Yao, Xiangyun Lu, Qiong Chen, et al, *Patient-derived mutations impact pathogenicity of SARS-CoV-2*, doi: <https://doi.org/10.1101/2020.04.14.20060160>.
16. KIRILLOVA O V, *Entropy concepts and DNA investigations*, SPhys Lett A , 2000· 274(5 – 6); 247 – 53.
17. KOSLICKI D, *Topological entropy of DNA sequences*, Bioinformatics, 2011, 27(8); 1061 – 7.
18. Jeffrey Z. J. Zheng, Christian H. H. Zheng, *A framework to express variant and invariant functional spaces for binary logic*, Modern Geometric Computing for Visualization, Springer-Verlag, 73-89, 1992.
19. Jeffrey Z.J. Zheng, Christian H.H. Zheng and Tosiyaasu L. Kunii, *A Framework of Variant Logic Construction for Cellular Automata*, Cellular Automata - Innovative Modeling for Science and Engineering, Dr. Alejandro Salcido (Ed.), InTech Press, 2011.
20. Jeffrey Zheng, *Variant Construction from Theoretical Foundation to Applications*, Springer, 2019, <https://www.springer.com/cn/book/9789811322815>
21. Jeffrey Zheng, Chris Zheng, Biometrics and Knowledge Management Information Systems, Chapter 11: Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019, 193-202 https://link.springer.com/chapter/10.1007/978-981-13-2282-2_11.
22. Jeffrey Z.J. Zheng, Christian H.H. Zheng and Tosiyaasu L. Kunii, *A Framework of Variant Logic Construction for Cellular Automata*, Cellular Automata - Innovative Modeling for Science and Engineering, Dr. Alejandro Salcido (Ed.), InTech Press, 2011.

Figures

Figure 1

Error bar diagram of topological entropy in six maps (a)-(f); (a)different countries; (b) USA; (c) China; (d) Australia;(e)Italy; (f)France

Figure 2
 scatter distribution of topological entropy in six maps (a)-(f); (a)different countries; (b) USA; (c) China; (d) Australia;(e)Italy; (f)France