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 β-coronavirus family. SARS-CoV-2 is highly similar to Pangolin CoV and BatCoV RaTG. Advanced scientific researches 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 mutation 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 bar 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 quantity based on C1 and C4 modules for visualizations. Sample results are shown that the method is powerful and useful to integrate all genomes on one unique genomic index map consistently. 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

Introduction
Severe acute respiratory syndrome coronavirus (SARS-CoV-2), originally temporarily named 2019-nCoV [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 β-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, bats maybe 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-SL-CoVZC45 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.

Citation: Jeffrey Zheng., et al. “Visualizations of Topologic Entropy on SARS-CoV-2 Genomes in Multiple Regions”. EC Neurology SI.02 (2021): 86-93.
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 that 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 BatCoVRaTG 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 S1 protein and RaTG13 [11]. The results of a recent study [12] have shown that the critical atomic interaction between the spurious 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-19 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 200000 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]. At present, it has been observed that novel coronavirus 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. There are also studies. 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 $C_1$ and $C_4$ output modules of 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 30000 bp. 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 2-4.

<table>
<thead>
<tr>
<th>Country</th>
<th>Number</th>
<th>Country</th>
<th>Number</th>
<th>Country</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belgium</td>
<td>218</td>
<td>China</td>
<td>200</td>
<td>Brazil</td>
<td>14</td>
</tr>
<tr>
<td>Italy</td>
<td>17</td>
<td>USA</td>
<td>675</td>
<td>Australia</td>
<td>60</td>
</tr>
<tr>
<td>France</td>
<td>26</td>
<td>Japan</td>
<td>92</td>
<td>Canada</td>
<td>35</td>
</tr>
</tbody>
</table>

*Table 1: Statistical table of virus sequences in different countries.*

**Citation:** Jeffrey Zheng., et al. “Visualizations of Topologic Entropy on SARS-CoV-2 Genomes in Multiple Regions”. EC Neurology SI.02 (2021): 86–93.
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 information 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] proposes to use 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 get an exponential growth rate, which is closest to the complexity of the sequence. Defined as:

\[
\lim_{n \to \infty} \frac{\log_2(P_{w^{2n}+n-1}(n))}{n}
\]

In the formula, \( w \) represents the sequence length, \( n \) is the sequence substring length, which satisfies \( 2^{2n+n-1} \leq |w| \leq 2^{2n+1}(n+1)-1 \). \( P_w^{2n}+n-1(n) \) 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 30000 bp and the calculated \( n \) value is 7 and the \( n \) value is 8 for the experiment.

Topological entropy generated

Citation: Jeffrey Zheng., et al. “Visualizations of Topologic Entropy on SARS-CoV-2 Genomes in Multiple Regions”. EC Neurology SI.02 (2021): 86-93.

<table>
<thead>
<tr>
<th>State/city</th>
<th>Number</th>
<th>State/city</th>
<th>Number</th>
<th>State/city</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>6</td>
<td>Arizona</td>
<td>5</td>
<td>California</td>
<td>11</td>
</tr>
<tr>
<td>Connecticut</td>
<td>8</td>
<td>Florida</td>
<td>9</td>
<td>Georgia</td>
<td>12</td>
</tr>
<tr>
<td>Illinois</td>
<td>6</td>
<td>Lowa</td>
<td>7</td>
<td>New Hampshire</td>
<td>4</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>69</td>
<td>Massachusetts</td>
<td>19</td>
<td>Minnesota</td>
<td>10</td>
</tr>
<tr>
<td>New York</td>
<td>188</td>
<td>Pennsylvania</td>
<td>6</td>
<td>South Carolina</td>
<td>7</td>
</tr>
<tr>
<td>Utah</td>
<td>15</td>
<td>Virginia</td>
<td>16</td>
<td>Washington</td>
<td>210</td>
</tr>
<tr>
<td>North Carolina</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>State/city</th>
<th>Number</th>
<th>State/city</th>
<th>Number</th>
<th>State/city</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>10</td>
<td>Chongqing</td>
<td>2</td>
<td>Foshan</td>
<td>3</td>
</tr>
<tr>
<td>Guangdong</td>
<td>8</td>
<td>Guangzhou</td>
<td>6</td>
<td>Hangzhou</td>
<td>37</td>
</tr>
<tr>
<td>Jiangsu</td>
<td>4</td>
<td>Nanchang</td>
<td>10</td>
<td>Shandong</td>
<td>3</td>
</tr>
<tr>
<td>Shanghai</td>
<td>42</td>
<td>Shangrao</td>
<td>8</td>
<td>Shenzhen</td>
<td>5</td>
</tr>
<tr>
<td>Wuhan</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3:** Statistical table of virus sequences in different regions of China.

<table>
<thead>
<tr>
<th>State/city</th>
<th>Number</th>
<th>Stat/eity</th>
<th>Number</th>
<th>State/city</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victoria</td>
<td>18</td>
<td>Abruzzo</td>
<td>2</td>
<td>Grand-Est</td>
<td>2</td>
</tr>
<tr>
<td>Queensland</td>
<td>16</td>
<td>Friuli Venezia Giulia</td>
<td>4</td>
<td>Hauts de France</td>
<td>9</td>
</tr>
<tr>
<td>Western Australia</td>
<td>9</td>
<td>Rome</td>
<td>6</td>
<td>Ile de France</td>
<td>15</td>
</tr>
<tr>
<td>Sydney</td>
<td>17</td>
<td>Veneto</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4:** Statistical tables of virus sequences in different regions of Australia, Italy and France.
Visualizations of Topologic Entropy on SARS-CoV-2 Genomes in Multiple Regions

Input: \( N \)

Output: \([\{C_1, C_2, C_3\} / \text{Connections}]\)

(Two real numbers, one \( \in [0, \log_2(m + 1)] \) and another one \( \in [0, 1] \))

Process: Selected one of \( C_1, C_2, C_3 \) entropy plus relevant connection lines

CF: 3 (Total number of selections).

\( \text{TE} \) is the \( C_4 \) part of four modules \( \{\text{CE, IE, ME, TE}\} \). The process of \( \text{TE} \) 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 get the topological entropy of the sequence [22].

**Result 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.

Figure 1a-1f 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 figure 1a and the actual data, it is not difficult to find that the average topological entropy of different countries is about \( H(w) = 0.9560 \), and the data topological entropy \( 0.9560 \) is likely to be an initial feature of SARS-CoV-2. In figure 1b, we can observe that the entropy of the genome in most parts of the United States is higher than \( 0.9560 \), corresponding to figure 1a, 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 figure 1c. 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. Figure 1d and 1e Australia and Italy also show similar characteristics to Wuhan, and the error range of the maximum average entropy is also the largest. In figure 1b, the United States shows the most significant margin of error in Illinois. However, in figure 1d, 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. Figure 1f 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 \( 1/10000 \).

Figure 2a-2f shows the scatter distribution of topological entropy indifferent countries or regions. \( n = 7 \) and \( n = 8 \) are selected as en-
Visualizations of Topologic Entropy on SARS-CoV-2 Genomes in Multiple Regions

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.

tropy calculation parameters, and the horizontal and vertical coordinates are entropy attributes.
In figure 2a, we can observe that some scattered points overlap and classify in different countries, such as the United States, Belgium, Australia, China's statistical aspects overlap and some mathematical points in the United States are separated. Next, specific to different cities in each country for observation. Comparing figure 2b with figure 2a, it is found that the separation of statistical points in the United States is related to the sequence of New York. In figure 2b, it can be clearly observed that the mathematical entropy points of New York are more characteristic than other locations. In figure 2c, we find that the entropy points of Wuhan are widely distributed, which mostly coincide with those of other cities, while the entropy points of Hangzhou are separated. Figure 2d shows the entropy scatter distribution in Australia. It can be observed that the sequence entropy values of Queen land and Victoria gather together, and the sequence entropy values of Sydney and Western Australia gather together. Figure 2e and 2f 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. Firstly, the average entropy value of each country is stable, and secondly, it is observed that there are differences in the entropy value of different cities. The entropy richness and entropy consistency of the virus 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 of Interest

No conflict of interest has been claimed.
Acknowledgements

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

Funding

This work was supported by NSF of China (62041213), the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018Z1002), NSF of China (61362014), Yunnan Advanced Overseas Scholar Project.

Bibliography

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


© All rights reserved by Jeffrey Zheng., et al.