2D Visual Analysis of SARS-CoV-2

Ruoxue Wu

Yunnan University

Mu Qiao

Yunnan University

Jeffrey Zheng (✉ conjugatelogic@yahoo.com)

Yunnan University

Research Article

Keywords: genomic index, combinatorial entropy, sequence phylogenetic, SARS-CoV-2, visualization

Posted Date: January 9th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-68275/v2>

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

[Read Full License](#)

2D Visual Analysis of SARS-CoV-2

Ruoxue Wu, Mu Qiao, Jeffrey Zheng

Abstract COVID-19 is outbreaking in worldwide. It caused millions of infections, killing hundreds of thousands of people and making all countries loss immeasurable trade. For finding the secret of SARS-CoV-2, researchers need to analyze various variation information such as multiple coronaviruses in different times over distinct countries. In this paper, the metagenetic analysis system MAS is used to analyze SARS-CoV-2 genomes collected from different countries as input datasets, and special genomic indices are provided to be a global characteristic quantity based on the A1 and C1 modules of the MAS for visualizations. In this method, one RNA sequence is split into M segments and counting the number of genetic probability measures for 16 combinations of four genomic symbols. After statistical probability processes, each probability distribution can be transferred into an entropy quantity on both 2D and 1D histograms to show these results for all collected genomes. Under this approach, a pair of combinatorial entropies determine a 2D genomic index map to generate a heatmap for more massive clusters of genomes with similarity contents to provide basic quantitative in variants to organize further collected genomes as a construction of a phylogenetic tree. Further explorations are required. **Keyword** genomic index, combinatorial entropy, sequence phylogenetic, SARS-CoV-2, visualization

Ruoxue Wu¹ e-mail: 691068796@qq.com
Mu Qiao^{2,3} e-mail: qiaomu@mail.ynu.edu.cn
Jeffrey Zheng^{2,3,4} e-mail: conjugatelogic@yahoo.com

¹ Beijing

² Yunnan University, Kunming

³ Key Laboratory of Software Engineering of Yunnan

⁴ Key Laboratory of Quantum Information of Yunnan

This work was supported by the NSFC (62041213), the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018ZL002), the NSFC (61362014), Yunnan Advanced Overseas Scholar Project.

Introduction

Today all the world is facing the most significant challenge after 1918. COVID-19 is sweeping the world in three months. There are about 85,000 patients in China and 3 million cases in the world. Mainly, the United States is the worst affected country, which has nearly 1 million patients of SARS-CoV-2. Moreover, this disease has a higher fatality rate, which can achieve about 10%. Therefore, to find the treatment and antidotes, we need to analyze the coronavirus.

As we know, the coronavirus is composed of RNA. Due to the instability of the RNA construct, the RNA virus is more accessible to evolution and mutation. According to [1], there are five subtypes of SARS-CoV-2 in the world. Some countries have one subtype, and some countries have two or more subtypes. So comparing these RNA sequences is very important. We must find the same or different points to distinguish different subtypes of the coronavirus.

In this paper, we utilize the metagenetic analysis system MAS especially on the C1 and A1 modules to process the RNA sequence of coronavirus in different countries. Besides, we analyze these results by graphs.

Method

In this section, we will introduce our method and its principle.

Pre-process

According to [2], we need to pre-process the RNA sequence of coronavirus. The whole sequence is divided into several parts. The specific steps are as follows. We get the data of the coronavirus RNA sequence from the GISAID database at the beginning. Next, we segment the sequence N into several subsequences whose length is m . So there are $M = N/m$ subsequences[2]. We will analyze these subsequences to get the information about the type of coronavirus.

Variant Construction for SARS-CoV-2

First, we should get the subsequence of the coronavirus's sequence, use variant construction to analyze each subsequence. Based on the[3], we account for the number of bases and their combinations in each subsequence. We can use Equation (1) to get the result.

$$p_b = \frac{m_b}{m}, b \in \{A, C, G, T\} \quad (1)$$

where m_b is the number of bases and their combinations in the subsequence [4]. So there are 15 untrivial combinations like A, T, C, G, A+T, A+C, A+G, T+C, T+G, C+G, A+T+C, A+T+G, A+C+G, T+C+G, A+T+C+G.

Entropy

Entropy is an efficient method to get the targets' information. We calculate the entropy of coronavirus' RNA sequence. Based on the previous section, we measure the M subsequences' entropy to achieve information on distributions on vectors or matrices. According to Section 2.2. we can know that there are $m + 1$ vectors or $(m + 1)^2$ matrices. The entropy $E(X)$ can be measured by Equation(2).

$$E(X) = - \sum_{j=0}^m p_j \log_2(p_j) \quad (2)$$

where $E(X) \in \mathbf{R}$ and also is restricted in $[0, \log_2 T]$, p_j represents the j^{th} probability which can be get by $\frac{Z_j}{m}$, $0 \leq Z_j \leq m$, and Z donates that a matrix or vector which has $m + 1$ or $(m + 1)^2$. So T is $m + 1$ or $(m + 1)^2$.

Visualization

After getting each subsequence's entropies, we count the same value of these entropies and use 2D coordinate to represent the results. X -axis coordinate means that the value of the entropy. Y -axis coordinate is the total number. Finally, we can get a histogram like Figure 1.

Fig. 1 The entropy histogram of Wuhan sample

Besides, we also choose two combinations to form the 2D scattergram's X - and Y -axis coordinates based on different coronavirus sequences. The example is shown in Figure 2.

Fig. 2 The 2D scattergram of different sequences in six different countries

Sequence Phylogenetic

Based on [5], we can know that due to RNA characteristic, the coronavirus is evolving continually. So For better understanding SARS-CoV-2, we also estimate the evolution time. According to [2], we can know that each sequence has a unique entropy. So we calculate the difference value of each entropy like Equation (3) and combine the time in sequence's data to analyze the sequence phylogenetically.

$$S = |E_a - E_b| \quad (3)$$

In Eq.3, S means the difference value, E_a and E_b represent the entropy value of two different sequences respectively[6].

Results

In this section, we will show the visual result of our method for the SARS-CoV-2.

Conclusion

In this paper, we analyze the types of SARS-CoV-2 in different countries by the metagenetic analysis system and their entropies. In order to show these results

(a) The A+T bases in samples of Australia.

(b) The A+T bases in samples of Brazil.

(c) The A+T bases in samples of United States.

(d) The A+T bases in samples of Britain.

(e) The A+T bases in samples of China.

Fig. 3 (A+T)% in multiple segments

(a) The entropy result samples of Australia.

(b) The entropy result samples of Italy.

(c) The entropy result samples of United States.

(d) The entropy result samples of Britain.

(e) The entropy result samples of China.

(f) The entropy result samples of France.

Fig. 4 The entropy result in multiple segments

clearly and efficiently, we use a histogram to project these results. Also, getting the evolution time of different coronavirus' types, we use entropy to get the result.

Conflict Interest

No conflict of interest has been claimed.

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

References

1. Peter Forster, Lucy Forster, Colin Renfrew, and Michael Forster, "Phylogenetic network analysis of sars-cov-2 genomes," *Proceedings of the National Academy of Sciences*, vol. 117, no. 17, pp. 92419243, 2020.
2. Jeffrey Zheng, *Variant Construction from Theoretical Foundation to Applications*, Springer, 2019, <https://www.springer.com/cn/book/9789811322815>
3. Jeffrey Zheng and Christian H. Zheng, "A framework to express variant and invariant functional spaces for binary logic," *Frontiers of Electrical and Electronic Engineering in China*, vol. 5, no. 2, pp. 163172, Jun 2010.
4. 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.
5. Jeffrey Zheng, Christian H.H. Zheng and Tosiya 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.
6. L. Shanguan and K. Jamieson, "The design and implementation of a mobile RFID tag sorting robot," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, New York, NY, USA, 2016, MobiSys '16, pp. 3142, ACM.

Figures

Figure 1

The entropy histogram of Wuhan sample

Figure 2

The 2D scattergram of different sequences in six different countries

(a) The A+T bases in samples of Australia.

(b) The A+T bases in samples of Brazil.

(c) The A+T bases in samples of United States.

(d) The A+T bases in samples of Britain.

(e) The A+T bases in samples of China.

Figure 3

(A+T)% in multiple segments

(a) The entropy result samples of Australia.

(b) The entropy result samples of Italy.

(c) The entropy result samples of United States.

(d) The entropy result samples of Britain.

(e) The entropy result samples of China.

(f) The entropy result samples of France.

Figure 4

The entropy result in multiple segments