Similarity Comparison of Multiple Coronavirus Sequences from 2D to 1D Linearizing Transformation

Feng Deng, Jeffrey Zheng

Abstract Many studies on COVID-19 have been carried out, and it is interesting to apply methods and models to process the whole sequence of RNA. Similarity comparison of SARS-CoV-2 genomes plays a key role in naturally tracing its origin in scientific exploration, and further explorations are required. In this paper, an innovative of transformation from a 2D density matrix to 1D measuring vector is proposed based on the A5 module of the MAS for visualization. The core transformation projects whole RNA sequences of multiple coronaviruses in 2D matrices and then forms 1D measuring vectors on variant maps. The relationships of SARS-CoV-2 genomes are compared by their similarity properties and genomic index of entropy quantities applied to classify relevant results into groups.
Keywords: SARS-CoV-2, RNA sequence, density matrix, vector, 2D to 1D linearized, visualization

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

Feng Deng
Yunnan University, Kunming, e-mail: 1345246776@qq.com ·
This work was supported by the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018ZI002).

Introduction

Since the outbreak of COVID-19 in Wuhan, China, in December 2019, the epidemic has now been more than four months, and more than 100 countries in the world have been infected successively. According to the World Health Organization (WHO) situation report on April 22, 2020, the cumulative number of global diagnoses is 471136, while the death toll was 169006 because of this epidemic [1]. The study of SARS-CoV-2 from genomics has certain help for the origin and evolution, development and spread of diseases, clinical diagnosis and treatment, antiviral emergency drugs and antibody drugs [2]-[6]. Thousands of SARS-CoV-2 genomes from many countries can be found on the website GISAID. Homology modeling is mainly used to explore the possible receptor binding characteristics of viruses, and it is the main method for comparing gene sequence similarities. An existing study compared SARS-CoV-2 sequences from 6 patients in Wuhan with SARS and MERS sequences [7]. One study used 9 gene sequences and found that SARS-CoV-2 is similar to SARS [8]. Another study used only 5 sequences and found that SARS-CoV-2 is similar to SARS [9]. In addition, the large sequence data analysis tool I-MLCS and similar algorithms are used in one paper to compare similarities between sequences [10]. Existing research lacks the exploration and research of the entire RNA sequence of SARS-CoV-2; therefore, it is also a worth thinking question that uses the whole sequence similarity to compare between viruses [12].

This paper proposes using the PMLP-V based on a variant system [13] to process the entire gene sequence. This visualization method is an innovative of transformation from a 2D density matrix to a 1D measuring vector and is based on the A5 module of the MAS. As an emerging technology method, its main idea is to use the 4-ary symbol as a meta-structure to deal with random sequences from cryptographic, DNA / RNA to ECG signals [14] and observe the global statistical distribution of sequences from an overall perspective feature. For the PMLP part, the basic mode starts with sequence input and ends with 1D variant map output. From variant map, the relationships of SARS-CoV-2 genomes are compared by their similarity properties. Finally, information entropy is used to demonstrate the results of variant maps and to classify relevant results into groups.

Model and Methods

PMLP-V

Using the variant system that includes three major theories: variant logic, variant measurement, and variant map [14], in the field of big data is an innovative method of thinking research, and this variant construction has a good expression in sequence processing.

The processing of RNA sequences based on variant logic consists of three main parts: sequence inputting, module processing (PMLP), 1D diagram outputting and verification. The basic framework is shown in Fig 1.

PMLP: Processing，Measurement，Linearization，Projection.

V: Verification.

Processing: Enter any one virus sequence into the program. In the processing module, a fixed length k is used to divide the whole sequence into several segments. In the measurement module, one selected segment sequence is used as one unit to count the number of bases. The base combination AT, AC is chosen as the position coordinate with a value of 1 at this point. If the value of this coordinate is other values not 1, add 1 to the existing value. The 2D density matrix is output after traversing all the number of sequence segments. In the linearization module, a linearized matrix is obtained by transforming the 2D density matrix to a 1D measuring vector. Then, the measuring vector is projected to be a 1D variant map in the projection module. Finally, verify the results and try to classify relevant results into groups.

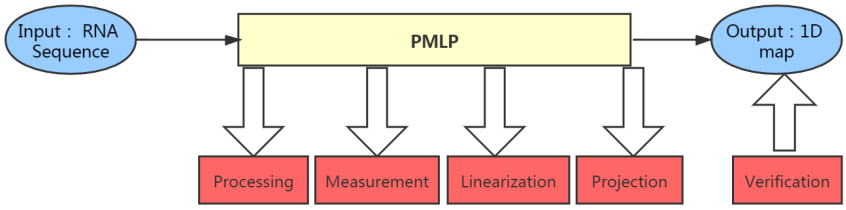

Fig. 1 Basic Framework

A. Processing

The main job of the processing module is to segment the entire RNA sequence to ensure that the length of each processed subsequence is the same.

B. Measurement

The measurement module mainly obtains a 2D density matrix by counting the number of bases of each subsequence. Use the number of bases as the horizontal and vertical coordinates to construct a density matrix. The main statistics in this article are the number of a pair of base combinations AT, AC, with (num_{AT}, num_{AC}) as the row and column of the density matrix. If position is first occurrence, record as 1. If it appears multiple times, add 1 to the original value.

C. Linearization

The main function of linearization is to transform from a 2D density matrix to a 1D measuring vector by retaining valid values and deleting all 0 elements. After performing this operation, output a one-dimensional matrix.

D. Projection

RNA sequence visualization. In this module, the whole sequence is projected to a 1D variant map.

E. Verification

The main work in this module is to utilize information entropy to verify the map results and classify them.

Details

Processing

N: Number of files.

M^q:Length of whole RNA base sequence of the q-th virus file.

k: Segment length.

D^q: Number of subsequences.

$$D^q = \frac{M^q}{k} \qquad q \in N \tag{1}$$

Calculating

Locate: 2D density matrix.

Entropy: Information entropy.

num_A^q (i): Number of A in the i-th sequence of the q-th virus file.

num_T^q (i): Number of T in the i-th sequence of the q-th virus file.

num_C^q(i): Number of C in the i-th sequence of the q-th virus file.

num_{AT}^q (i): Number of AT in the i-th sequence of the q-th virus file.

num_{AC}^q (j):Number of AC in the j-th sequence of the q-th virus file.

$P_{AT}q^q$(t): Probability of AC in the t-th sequence of the q-th virus file.

$$num_{AT}^q(i) = num_A^q(i) + num_T^q(i) \qquad q \in N, \qquad i \in D^q \tag{2}$$

$$num_{AC}^q(j) = num_A^q(j) + num_C^q(j) \qquad q \in N, \qquad j \in D^q \tag{3}$$

$$Locate[num_{AT}^q(i), num_{AC}^q(i)]_{=} Locate[num_{AT}^q(i), num_{AC}^q(i)] + 1 \quad q \in N,\ i \in D^q \tag{4}$$

$$Entropy = \left(-\sum_t^{D^q} P_{AT}^q(t) \log_2 \left(P_{AT}^q(t) \right) \right) / D^q \quad t \in D^q \tag{5}$$

Results and Analyses

Nine typical virus sequence files named China-COVID-19, HCov-HKU1, HCoV-NL63, Pangolin, HCov-OC43, MERS,SRAS, Ebola and USA-COVID-19 were selected in the article.

SARS-CoV-2: China-COVID-19 and USA-COVID-19.

Influenza Coronavirus: HCov-HKU1, HCoV-NL63 and HCov-OC43.

Pangolin: Pangolin.

Highly pathogenic and deadly: MERS, SRAS and Ebola .

Variant Maps

Figs. 2-4 show the results of different parameters $(k = 2, 4, 8)$. Ordinate is the number of statistical projections while abscissa indicates the number of valid values in the position matrix.

To effectively distinguish similarity of the viral sequences, it is recommended to select a smaller k-value.

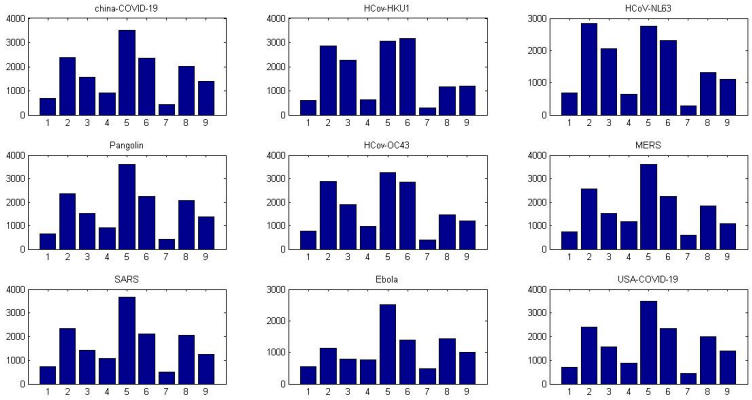

Fig. 2 $k = 2$

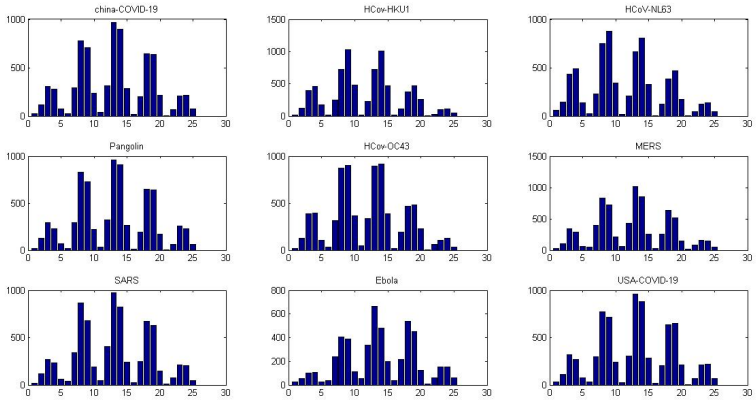

Fig. 3 $k = 4$

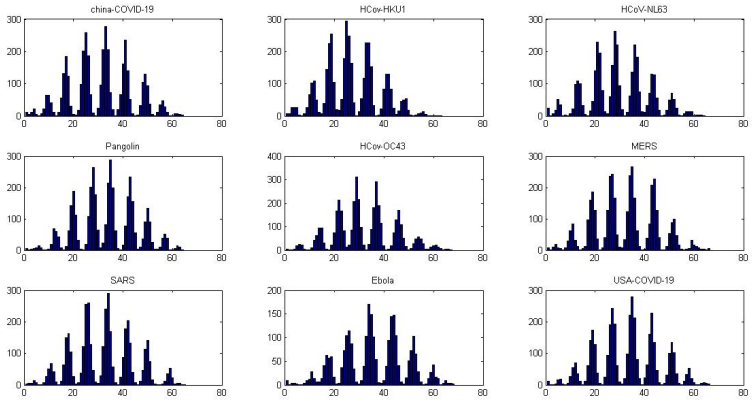

Fig. 4 $k = 8$

Entropy curves

Information entropy can be used as a measure of judging system complexity. The system is more complex, and the entropy is larger.

Each curve corresponds to a sequence. The ordinate represents the average information entropy, and the abscissa represents the value of fixed parameters,which are k=2, $2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8$. Each parameter corresponds to an average information entropy. That is, a complete viral RNA sequence corresponds to 8 average information entropies, and then fits them into a curve and outputs it.

Analyses

Figs 2-4: Judging similarity based on the distribution of histograms between variant maps, that is, if maps are intuitively the same, the two viruses are considered similar. Taking the sequence of China-COVID-19 file as a benchmark, virus that is the best similar to China-COVID-19 was selected.

Fig 2: In this diagram, there are four viral sequences similar to China-COVID-19, which are Pangolin, MERS, SRAS and USA-COVID-19.

Fig 3: For this diagram, four viruses are similar: China-COVID-19, Pangolin, SRAS and USA-COVID-19.

Fig 4: In this diagram, Pangolin is the most similar to China-COVID-19 except USA-COVID-19.

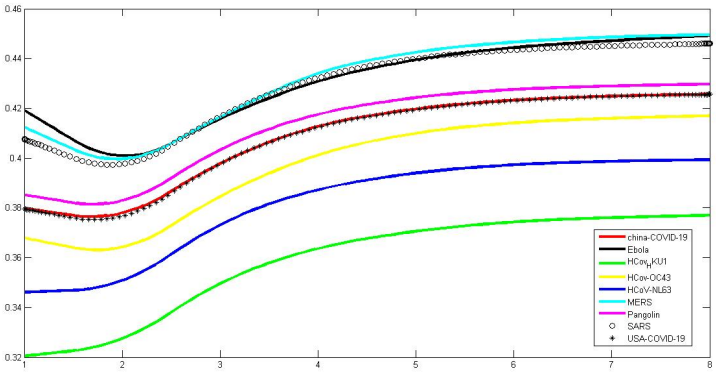

Fig. 5 9 curves of entropy

Fig 5: Comparing the curves of 9 sequences, it can be observed that the red line (China-COVID-19) and black star (USA-COVID-19) have the highest coincidence, indicating that the gene distribution of the two groups is visually similar. However, there is also a slight difference, and it is speculated that SARS-CoV-2 exhibits gene recombination and mutation. Except for USA-COVID-19, the difference between Pangolin (purple line) and China-COVID-19 is the smallest, indicating that there is not much difference in the proportion of bases between them. The internal complexity of the systems are similar, as are the gene sequences.

Final result: As the analyses aboove show, SARS-CoV-2 is similar to Pangolin and may belong to a homologous sequence.

Conclusion

This paper proposes using a variant logic system to process virus genomes, and transforms RNA data to a 1D variant map. Utilizing visual analysis and special transformation methods to compare genome similarity is the main idea. Finally, we demonstrate the comparison results by using information entropy curves. The analysis results show that SARS-CoV-2 genomes are highly similar to Pangolin virus, which is consistent with existing research results[10].

Variant logic has great advantages in processing big data. Its processing flow is simple, data loss is small and the output result is ideal. It provides a new idea for processing big data.

Conflict Interest

No conflict of interest has been claimed.

Acknowledgements

The authors would like to thank NCBI,GISAID, CNGBdb and Nextstrain for providing invaluable information on the newest dataset collections of SARS-CoV-2 and other coronavirus genomes.

References

1. *Coronavirus disease 2019 (COVID-19) Situation Report 93*. 2020.https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200422-sitrep-93-covid-19.pdf?$sfvrsn = 35cf80d7_4$.
2. Wu A, Peng Y, Huang B, et al. *Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China*. Cell Host Microbe. 2020 Mar 11;27(3):325-328. doi: 10.1016/j.chom.2020.02.001. Epub 2020 Feb 7. PMID: 32035028.
3. Zhang T, Wu Q, Zhang Z.*Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak*. Curr Biol. 2020 Mar 19. pii: S0960-9822(20)30360-2.doi: 10.1016/j.cub.2020.03.022. [Epub ahead of print] PMID: 32197085
4. Rehman SU, Shafique L, Ihsan A, Liu Q. *Evolutionary Trajectory for the Emergence of Novel Coronavirus SARS-CoV-2*. Pathogens. 2020 Mar 23;9(3). pii: E240.doi: 10.3390/pathogens9030240. PMID: 32210130
5. Rehman SU, Shafique L, Ihsan A, Liu Q.*Evolutionary Trajectory for the Emergence of Novel Coronavirus SARS-CoV-2*. Pathogens. 2020 Mar 23;9(3). pii: E240.doi: 10.3390/pathogens9030240. PMID: 32210130
6. K.G. Andersen, A. Rambaut, W.I. Lipkin, et al.*The proximal origin of SARS- CoV-2*. Nat,Med (2020).https://doi.org/10.1038/s41591-020-0820-9.
7. Xintian Xu, Ping Chen and et al, *Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission*.Sci. China Life Sci,Vol.63(457-460) 2020.https://doi.org/10.1007/s11427-020-1637-5.
8. Roujian Lu, Xiang Zhao and et al, *Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding*.The Lancet,542-545,2020.
9. Zhou P, Yang XL, Wang XG, et al. *A pneumonia outbreak associated with a new coronavirus of probable bat origin*. Nature. 2020 Mar;579(7798):270-273.doi: 10.1038/s41586-020-2012-7. Epub 2020 Feb 3. PMID: 32015507
10. Yanni Li , Bing Liu and et al, *Similarities and Evolutionary Relationships of COVID-19 and Related Viruses*. 2020.
11. R. Yan, Y. Zhang, Y. Li, et al. *Structural basis for the recognition of SARS-CoV- 2 by full-length human ACE2*. Science 27 Mar 2020: Vol. 367, Issue 6485, pp. 1444-1448.doi: 10.1126/science.abb2762.
12. Li C, Yang Y, Ren L. *Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species*. Infect Genet Evol. 2020 Mar 10;82:104285.doi: 10.1016/j.meegid.2020.104285. [Epub ahead of print] PMID: 32169673
13. Zheng, J.*Conditional Probability Statistical Distributions in Variant Measurement Simulations*. Acta Photonica,2011,40,1662-1666.https://doi.org/10.3788/gzxb20114011

14. Jeffrey Zheng, *Variant Construction from Theoretical Foundation to Applications*. Springer,2019.https://doi.org/10.1007/978-981-13-2282-2_1
15. Song Z, Xu Y, Bao L, et al.*From SARS to MERS, thrusting coronaviruses into the spotlight*.Viruses. 2019; 11: 59.https://doi.org/10.3390/v11010059