2D Similarity Map of Multiple Coronavirus Gene Sequences

Huaxian Zheng, Jeffrey Zheng

Abstract The outbreak of a novel coronavirus (SARS-CoV-2) in many countries in the world from late 2019 to 2020 resulted in millions of infected people, and caused serious damage to the social environments with significant changes in human power and material resources in the world. The novel coronavirus is an RNA virus. RNA mutation is common in nature. This makes it extremely difficult to develop a virus vaccine in a short period. The evolution of the virus has been in a mutation state, in which a certain sequence changes associated with time and environments in similar distributions. A larger number of genomes were collected in various open source databases for scientists in further explorations. In this paper, a 2D similarity comparison scheme on the A2 module of the MAS is proposed for extracting internal information among a genome undertaken M segment partitions to provide visual results based on probability measures and quantitative statistics. First, a genome is segmented into corresponding numerical transformations, and then four numbers of meta symbols in each segment are counted. Corresponding probability measures are calculated. Second, the probability is transformed into polar coordinates, and the polar coordinates are mapped into a $M \times M$ matrix. Then, a 1D genome can be processed into 2D measures with similarity properties in sequence. Through this correlation matrix, relevant similarity results are analyzed.

Keyword coronavirus genome, 1D sequence, 2D matrix, diagram, probability measure, statistics, 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

Huaxian
Yunnan University, Kunming, e-mail: 12017001024@ynu.edu.cn

This work was supported by the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018ZJ002), NSF of China (61362014), Yunnan Advanced Overseas Scholar Project.

Introduction

To date, novel coronavirus pneumonia (SARS-CoV-2) has been diagnosed in more than approximately three million cases according to the latest data. SARS-CoV-2 was identified as the seventh human coronavirus, and the other six were HCoV-229e, HCoV-NL63, HCoV-oc43, HCoV-hku1, severe acute respiratory syndrome (SARS-CoV) and respiratory syndrome coronavirus in the Middle East (MERS-CoV).

Among them, HCoV-229e, HCoV-NL63, HCoV-oc43 and HCoV-hku1 are prevalent in the global population, accounting for approximately one third of the common cold infections in the blood-brain barrier. Because it is difficult for most coronaviruses to judge the difference of viruses through the surface, they are judged by genome characteristics.

The traditional methods of sequence analysis include sequence alignment algorithms, for example, the algorithms based on dynamic programming to calculate DNA sequence similarity. With the rapid development of sequencing technology, a large number of data sets make the sequence analysis method based on alignment appear bottleneck in the calculation, so the following is the incubation of some non-alignment biological sequence analysis methods. The non-sequence alignment algorithm can be divided into statistical model-based methods, geometry-based methods and complexity-based sequence comparison methods. The commonly used methods to measure the relationship between DNA sequences are, the Markov model and Kullback Leibler difference (KLD), Mahalanobis distance, chaos theory, Euclidean distance, cosine distance, etc. In addition, Chunting Zhang uses the knowledge of geometry to analyze DNA sequences and proposes the Z-curve theory, which is an equivalent representation form of genome sequences. All the information of genome sequence information can be expressed on the Z-curve.

Novel coronavirus has been a pressing matter of the moment in novel coronavirus. It is imperative to determine the source of the outbreak and take measures to analyze the virus. The comparative analysis of the sequence of the virus in each country helps to analyze the coronavirus in all directions. The development of sequencing technology makes it easier to obtain a new coronavirus sequence. The international open source database GISAID has been keeping up with the new databases. A large number of coronavirus sequences submitted by various countries are not easy to observe directly, and the introduction of visualization technology can more intuitively analyze the sequences. For many years, there have been many visualization methods in the field of bioinformatics, including sequence comparative analysis, phylogenetic tree construction, database search, structure comparative analysis, and gene recognition.

The novel coronavirus base nucleic acid sequence is added every day in the database. It is unlikely that the structure of each sequence will be compared and analyzed one by one. Therefore, we must analyze and select the data accumulated by the computer in order to achieve the result that should be judged by experiment. At present, gene sequence visualization technology has been used in the fields of gene sequence similarity analysis, protein structure analysis and family tree construction. Therefore, it is necessary to study the method of gene sequence visualization in the

field of bioinformatics. In this paper, a 2D similarity comparison scheme on the A2 module of the MAS is proposed, which shows that different kinds of coronaviruses have their own distinct characteristics.

Model Description

In this paper, the visualization method is combined with the probability statistics method. First, the gene sequence is segmented to count the probability of each segment, then the probability is normalized to map the matrix, and then the graph is drawn by way of a thermodynamic diagram. The specific methods are as follows.

A set of sequences with values between $[0, 2]$ is obtained and visualized by using the Gracian angular field method. First, a time correlation matrix is created for each object (X_i, X_j) . Second, it rescales the time series in a range $[a, b]$, and $(-1 \leq a \leq b \leq 1)$. It then calculates the polar coordinates of the scaled time series. Finally, the sine of the angle difference between the angle and cosine of the Gramian angle, and field (Gasf) and the angle difference of the Gramian angle difference field (Gadf) are calculated. The steps are as follows.

First, scale the sequences S1, S2, S3, S4 to between $[-1, 1]$ by formula (1).

$$\tilde{X}_i = a + (b - a) * \frac{X_i - \min(x)}{\max(x) - \min(x)}, \forall i \in \{1, \dots, n\} \quad (1)$$

Second, the new sequence $\tilde{X}_i$ in polar coordinates can display the formula of the encoded angle cosine such as (2). The data recalibrated in different intervals have different angle bounds $[0, 1]$ corresponding to the cosine function value in $[0, \pi/2]$, $[-1, 1]$ mapping to the cosine value between $[0, \pi]$.

$$\Theta_i = \arccos(\tilde{X}_i), \forall i \in \{1, \dots, n\} \quad (2)$$

Finally, after the rescaled time series is transformed into a polar coordinate system, the time correlation in different time intervals can be identified by considering the triangle and/or difference between each point, and the angle perspective is used. GASFs are defined in equations (3)-(10).

$$GASF_{i,j} = \cos(\theta_i + \theta_j), \forall i,j \in \{1, \dots, n\} \quad (3)$$

$$GADF_{i,j} = \sin(\theta_i - \theta_j), \forall i,j \in \{1, \dots, n\} \quad (4)$$

$$GASF = \cos(\theta_i + \theta_j) = \tilde{X} \cdot \tilde{Y} - \sqrt{I - \tilde{X}^2} \cdot \sqrt{I - \tilde{Y}^2} \quad (5)$$

$$GASF = \sin(\theta_i - \theta_j) = \tilde{X} \cdot \sqrt{I - \tilde{Y}^2} - \tilde{Y} \cdot \sqrt{I - \tilde{X}^2} \quad (6)$$

The calculation formula of Gasf is (7):

$$G = \begin{bmatrix} \cos(\theta_1 + \theta_1) & \dots & \cos(\theta_1 + \theta_n) \\ \cos(\theta_2 + \theta_1) & \dots & \cos(\theta_2 + \theta_n) \\ \vdots & \ddots & \vdots \\ \cos(\theta_n + \theta_1) & \dots & \cos(\theta_n + \theta_n) \end{bmatrix} \quad (7)$$

$$\cos(\theta_1 + \theta_2) = \cos(\arccos(x) + \arccos(y)) \quad (8)$$

$$= \cos(\arccos(x)) \cdot \cos(\arccos(y)) - \sin(\arccos(x)) \cdot \sin(\arccos(y)) \quad (9)$$

$$= x \cdot y + \sqrt{1-x^2} \cdot \sqrt{1-y^2} \quad (10)$$

The mapping matrix used in this paper is similar to the Gram matrix, $v_1, \dots, v_n$ in linear algebra. The Gram matrix of a family of vectors of inner product space is the symmetry of the inner product, and its element is $G_{i,j} = (v_j | v_i)$. Gram matrix theorem: let X be an inner product space, $u_1, u_2, \dots, u_n \in X$ matrix:

$$G = \begin{bmatrix} (u_1, u_1) & (u_2, u_1) & \dots & (u_n, u_1) \\ (u_1, u_2) & (u_2, u_2) & \dots & (u_n, u_2) \\ \vdots & \vdots & \ddots & \vdots \\ (u_1, u_n) & (u_2, u_n) & \dots & (u_n, u_n) \end{bmatrix} \quad (11)$$

Data processing with gram-like corner field:

- A. Get data s
- B. Compress its data range to $[-1, 1]$
- C. Selection method sin or cos processing
- D. Then use the piecewise aggregation approximation (PAA) to reduce the size of each time series
- E. The image is flattened to one dimension

Visual Results

Typical human coronaviruses are shown in Table 1.

In the visualization model, the change of parameter value will have a great influence on the result of graphics. Whether the parameter value is too large or too small, some distribution features of the image will be hidden. The main two parameters in this paper are segment value m and sliding window value X . Six groups of novel coronavirus sequences were selected for comparative analysis. The m values were the same, and the X values showed different figures as shown in Figure 1. The X values were the same, and the m values were different in Figures 2(a1)-(a3) and (b1)-(b3) were selected as the 2019-nCoV MN908947 and the human coronavirus NL63 MN908947 respectively.

Table 1 Sequence of human coronaviruses

Sample	No.	Collection years	Host	Size
HCoV-229E	NC_002645	1966	Bats	28KB
HCoV-NL63	NC_005831	2004	Palm Civets,Bats	28KB
HCoV-OC43	NC_006213	1967	Cattle	31KB
SARS-CoV	AY485277	2003	Mice	30KB
HcoV-HKU1	NC_006577	2005	Palm Civets,Bats	31KB
MERS-CoV	JX869059	2012	Bats - Camels	31KB
SARS-CoV-2	MN908947	2019		31KB

Fig. 1 Comparison of different sliding window values with the same segment values

The corresponding characteristics of the RNA sequence are A, G, C, T. Its subsets are A, G, C, T, AG, AC, AT, GC, GT, CT, AGC, AGT, ACT, GCT, AGCT. Figure 3 shows the visualization results of the 15 subsets.

Fig. 2 Comparison of different segment values and the same sliding window values

Fig. 3 Fifteen projections of 2D similarity maps for SARS-CoV-2 sample

In the study of visualization of 15 seed set sequences of sequences, it is found that the features displayed by the subsets of the T-sequences are effective after many comparisons. Therefore, the 9 coronavirus sequences they are all treated as visualization diagrams of T-subsets, and Figure 4 is obtained. Compared with other sequence characteristics, both human coronavirus NL63 and 2019-nCoV MN908947 on the T projections are shown in Figure 5.

Fig. 4 Comparison of nine coronaviruses

Fig. 5 2D maps on the T-projections between NL63 and 2019-nCoV

The standard difference calculation is introduced to prove that the similar distribution observed by the naked eye is reflected in the difference. The standard difference is represented by a graph. Because the standard difference is 0 when the same sequence is compared, the part with the value of 0 is replaced by the mean value of the column in which it is located without affecting the change trend. Therefore, the scatter diagram between the standard differences of 9 sequences is shown in Figure 6. The information entropy of the sequence can reflect the degree of confusion between the sequences, and the larger the image information entropy is, the greater the amount of information. Figure 7 is a broken line diagram of sequence information entropy and graphical information entropy of 9 coronaviruses.

Fig. 6 scatter diagram of standard deviation of nine coronavirus sequences

Fig. 7 information entropy

Result Analysis

The change in the parameter value has a great influence on the graph results. In this paper, the gene sequence is visualized by controlling the change in segment value, window value and coordinate value, and its distribution characteristics are displayed one by one in the graph. In the case of the same control parameters, comparing different coronavirus sequences, it was found that the more similar the gene sequence in the family tree of the genome, the more obvious the similarity of its graphical distribution characteristics.

Phylogenetic analysis of novel coronaviruses, the RNA-dependent RNA polymerase (RdRp) gene and the S gene sequence indicated that RaTG13 is a close cousin of the new coronavirus and has different lineages than other SARSr-CoV. The novel coronavirus is highly similar to RaTG13 in the whole genome. Its complete genome sequence is 96.2% homologous to that of the SARS-like coronavirus isolated from China, and the similarity of nucleotide similarity to bat-SL covzc45 (GenBank MG772933) is 89.1%. The distribution of the novel coronavirus bases is similar to that of the SARS family. These similarities are shown in the pictorial distribution of the new coronavirus.

In many published research results, it has also been shown that the sequence of RaTG13 coronavirus collected from bats is highly similar to the sequence of new coronary disease. From Figure 4, it is also observed that the distribution of base characteristics of the two virus sequences is indeed highly similar to that of other coronavirus sequences. From the information entropy of the graph, it can be seen

that the entropy values of the sequence are similar. From the information entropy of the image, it can be observed that when the value is the T sequence, the entropy value of NL63 sequence is the largest, indicating that the amount of information obtained is also the largest.

Conclusion

The visual graph of this paper shows that the distribution of bases in different coronaviruses has their own characteristics, which is convenient for the analysis of virus sequences. In the research of this paper, the focus is to process several typical human coronavirus sequences. The nine selected coronavirus sequences are of great value in the field of coronavirus research. The data come from the NCBI database. The processing mode is to preprocess first and then use the variant visualization model to process the two-dimensional matrix. The first is to map the matrix for intuitive comparative analysis, and the second is to map the matrix into a graph that is available when measuring the similarity between sequences. Then the standard deviation is introduced to measure the base distribution of the sequence, and the matrix value subtraction can be used to calculate the standard deviation of the sequence. In this paper, the graphical results include a graphical comparison of the changes in the control parameters and a graphical comparison of the same parameters on different virus sequences.

In the visual results, the common distribution characteristics of virus sequences and the characteristics of their existence are observed, while the distribution of their base sequences among the same type of virus sequences has certain similarity. In the standard difference measurement, the distribution of the original sequences can be compared one by one. Correspondingly, the model in this paper is feasible.

Conflict Interest

No conflict of interest has been claimed.

References

1. Vijgen L , Keyaerts E , Moes E , et al. Development of One-Step, Real-Time, Quantitative Reverse Transcriptase PCR Assays for Absolute Quantitation of Human Coronaviruses OC43 and 229E, *Journal of Clinical Microbiology*, 43(11):5452-5456, 2005.
2. McBride R , Fielding B . The Role of Severe Acute Respiratory Syndrome (SARS)-Coronavirus Accessory Proteins in Virus Pathogenesis, *Viruses*, 4(12):2902-2923, 2012 .

3. Jones B A , Grace D , Kock R , et al. zoonosis emergence linked to agricultural intensification and environmental change *Proceedings of the National Academy of Sciences* , 110(21):8399-8404, 2013.
4. Hoek L V D .Human coronaviruses: What do they cause?, *Antiviral therapy*, 12(4 Pt B):651-658, 2007.
5. Arbour N , Day R , Newcombe J , et al. Neuroinvasion by Human Respiratory Coronaviruses, *Journal of Virology*, 74(19):8913-8921, 2000.
6. Jeffrey Z.J. Zheng, Christian H.H. Zheng and Toshiyasu 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.
7. Zhou Peng, Yang Xing-Lou, Wang Xian-Guang, Hu Ben, Zhang Lei, Zhang Wei, Si Hao-Rui, Zhu Yan, Li Bei, Huang Chao-Lin, Chen Hui-Dong, Chen Jing, Luo Yun, Guo Hua, Jiang Ren-Di, Liu Mei-Qin, Chen Ying, Shen Xu-Rui, Wang Xi, Zheng Xiao-Shuang, Zhao Kai, Chen Quan-Jiao, Deng Fei, Liu Lin-Lin, Yan Bing, Zhan Fa-Xian, Wang Yan-Yi, Xiao Geng-Fu, Shi Zheng-Li. A pneumonia outbreak associated with a new coronavirus of probable bat origin, *Nature*, 2020.
8. TTY Lam et al. Identification of 2019-nCoV related coronaviruses 1 in Malayan pangolins in southern China, *doi: <https://doi.org/10.1101/2020.02.13.945485>*, 2.18, 2020.
9. Mao, Y., Zheng, J. and Liu, W. (2017) Mapping Whole DNA Sequence on Variant Maps, *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 1037-1040, Sydney, 31 July-3 August 2017.
10. R. P. Stanley, *Enumerative Combinatorics*, Vol. 1, 2nd edition, Cambridge University Press, 1997.
11. D. Stanton, R. Stanton and D. White, *Constructive Combinatorics*, Springer-Verlag, 1986.
12. Y. H. Tang, *Binary Homogeneous Symmetric Polynomials and Binomial Theorem*, Zhejiang University Press, 2012 (in Chinese).
13. Li W , Moore M J , Vasilieva N , et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus, *nature*, 426(6965):450-454, 2003.
14. L. X. Wang, *An Elementary Treatise on Combinations*, Harbin Institute of Technology Press, 2012 (in Chinese).
15. Yeager C L , Ashmun R A , Williams R K , et al. Human aminopeptidase N is a receptor for human coronavirus 229E, *nature*, 357(6377):420-422, 1992.
16. Yu J F , Sun X , Wang J H . TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications, *Journal of Theoretical Biology*, 261(3):459-468, 2009.
17. Zhou Peng, Yang Xing-Lou, Wang Xian-Guang, Hu Ben, Zhang Lei, Zhang Wei, Si Hao-Rui, Zhu Yan, Li Bei, Huang Chao-Lin, Chen Hui-Dong, Chen Jing, Luo Yun, Guo Hua, Jiang Ren-Di, Liu Mei-Qin, Chen Ying, Shen Xu-Rui, Wang Xi, Zheng Xiao-Shuang, Zhao Kai, Chen Quan-Jiao, Deng Fei, Liu Lin-Lin, Yan Bing, Zhan Fa-Xian, Wang Yan-Yi, Xiao Geng-Fu, Shi Zheng-Li. A pneumonia outbreak associated with a new coronavirus of probable bat origin, *Nature*, 2020.
18. Jeffrey Z. J. Zheng, Christian H. H. Zheng, A framework to express variant and invariant functional spaces for binary logic, *Frontiers of Electrical and Electronic Engineering in China*, 5(2):163-172, Higher Educational Press and Springer-Verlag, 2010.
19. Jeffrey Zheng, Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019 <http://www.springer.com/in/book/9789811322815>.
20. Jeffrey Zheng, Variant Construction Theory and Applications, Vol. 1: Theoretical Foundation and Applications, Science Press 2020 (Chinese, Formal Publishing Soon).
21. N. J. A. Sloane, The On-Line Encyclopedia of Integer Sequences, <http://oeis.org>.
22. Jeffrey Zheng, Chris Zheng, Biometrics and Knowledge Management Information Systems, Chapter 11: Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019, 193-202 , http://link.springer.com/chapter/10.1007/978-981-13-2282-2_11.