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

Dr. Deng Feng¹ and Dr. Jeffrey Zheng²*

¹Yunnan University, Kunming, China
²Key Laboratory of Quantum Information of Yunnan, Key Laboratory of Software Engineering of Yunnan, Yunnan University, Kunming, China

*Corresponding Author: Dr. Jeffrey Zheng, Key Laboratory of Quantum Information of Yunnan, Key Laboratory of Software Engineering of Yunnan, Yunnan University, Kunming, China.

Received: August 18, 2020; Published: September 11, 2020

Abstract

Many studies on COVID-19 are carried out, 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 to trace its origin in scientific exploration naturally, further explorations are required. In this paper, an innovative of transformation from 2D density matrix to 1D measuring vector is proposed to be based on A5 module in MAS for visualization. The core transformation projects whole RNA sequences of multiple coronaviruses in 2D matrices then forming 1D measuring vectors on variant maps. The relationships of SARS-CoV-2 genomes are compared by their similarity properties and finally 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

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 diagnosis is 471,136 while death toll was 169,006 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 results 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 the SARS-CoV-2 sequences from 6 patients in Wuhan with SARS and MERS sequence [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 also found that SARS-CoV-2 is similar to SARS [9]. Besides, 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 researchs 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 to use PMLP-V basing on variant system [13] to process the entire gene sequence. This visualization method is an innovative of transformation from 2D density matrix to 1D measuring vector and bases on A5 module in 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 observes the global statistical distribution of sequences from an overall perspective feature. For PMLP part, the basic mode starts with sequence inputting and ends with 1D variant maps outputting. From variant maps, the relationships of SARS-CoV-2 genomes are compared by their similarity properties. At last of this method, use information entropy to demonstrate the results of variant maps and try 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 3 main parts, sequence inputting, module processing (PMLP), 1D diagram outputting and Verification. The basic framework is shown in figure 1:


Processing: Enter any one virus sequence into the program. In Processing module, use a fixed length k 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. Base combination AT,AC is chosen as the position coordinate which the value is 1 at this point. If the value of this coordinate is other values not 1, add 1 to the existing value. 2D density matrix is output after traversing all the number of sequence segments. In the Linearization module, linearized matrix is obtained by transforming 2D density matrix to 1D measuring vector. And 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 group.

Figure 1: Basic framework.

Processing

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

Measurement

The Measurement module mainly obtains 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 \((\text{num}_{AT}, \text{num}_{AC})\) 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.

**Linearization**

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

**Projection**

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

**Verification**

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

**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.

\[D_q = M_q \in N (1)\]

\[k\]

**Figure**

**Citation:** Dr. Deng Feng and Dr. Jeffrey Zheng. “Similarity Comparison of Multiple Coronavirus Sequences from 2D to 1D Linearizing Transformation”. EC Neurology SI.02 (2020): 50-56.
Results and Analysis

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


Influenza Coronavirus: HCoV-HKU1, HCoV-NL63 and HCoV-OC43. 1 Pangolin: Pangolin.

Highly pathogenic and deadly: MERS, SRAS and Ebola.

Variant maps

Figure 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.

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

Figure 2: k = 2.

Figure 2: k = 4.
Entropy curves

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

Each curve in corresponds to a sequence. The ordinate represents average information entropy, and the abscissa represents value of fixed parameters, which used 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 fit them into a curve and output it.

Analysis

Figure 2-4: Judging similarity based on the distribution of histograms between variant maps, that is, if maps are intuitively same, the two viruses are considered similar. Take the sequence of china-COVID-19 file as benchmark, select virus which is the best similar to china-COVID-19.

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

Figure 3: For this diagram, four viruses are similar, which are china-COVID-19, Pangolin, SRAS and USA-COVID-19.

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

Figure 5: Comparing the curves of 9 sequences, it can be observed that 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 exists gene re-combination and mutation. Except 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, so are the gene sequences.

Citation: Dr. Deng Feng and Dr. Jeffrey Zheng. “Similarity Comparison of Multiple Coronavirus Sequences from 2D to 1D Linearizing Transformation”. EC Neurology SI.02 (2020): 50-56.
Final result: As the analysis above, SARS-CoV-2 are similar to Pangolin and may belong to homologous sequence.

Conclusion

This paper proposes to use variant logic system to process virus genomes, transform RNA data to a 1D variant map. Utilizing visual analysis and special transformation way to compare genomes 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 is highly similar to Pangolin virus, which is consistent with existing research result [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.

Acknowledgements

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

Funding

This work was supported by the Key Project on Electric Information and Next Generation ITTechnology of Yunnan (2018ZJ002).

Bibliography


Citation: Dr. Deng Feng and Dr. Jeffrey Zheng. “Similarity Comparison of Multiple Coronavirus Sequences from 2D to 1D Linearizing Transformation”. EC Neurology SI.02 (2020): 50-56.
Similarity Comparison of Multiple Coronavirus Sequences from 2D to 1D Linearizing Transformation


© All rights reserved by Dr. Deng Feng and Dr. Jeffrey Zheng.