EC Neurology Special Issue - 2021

Visualizations of Multiple Probability Measures for SARS-CoV-2 Genomes

Tan Yao and Jeffrey Zheng*

School of Software, Yunnan University, Kunming, Key Laboratory of Quantum Information of Yunnan, Key Laboratory of Software Engineering of Yunnan, Engineering Research Center of Cyberspace of Yunnan, China

***Corresponding Author:** Jeffrey Zheng, School of Software, Yunnan University, Kunming, Key Laboratory of Quantum Information of Yunnan, Key Laboratory of Software Engineering of Yunnan, Engineering Research Center of Cyberspace of Yunnan, China.

Received: August 18, 2020; **Published:** January 25, 2021

Abstract

SARS-CoV-2 genomes are collected from various open source genomic banks. A set of SARS-CoV-2 genomes are selected for visualization under both the A3 and C1 modules of the metagenomic analysis system MAS. Multiple probability measures are mapped as relevant 1D histograms, and it is convenient to observe distinct differences among various distributions to organize similar patterns into relevant groups. Sample genomes were processed and their visual results were illustrated.

Keywords: *Metagenomic Analysis System MAS; Variant Measurement; Variant Map; Multiple Probability; Normalized; 1D Histogram*

Introduction

Everyone is concerned about the outbreak caused by SARS-CoV-2 [6] and is making their own efforts to overcome this outbreak. In particular, the majority of medical personnel not only race against time but also fight against diseases. From the frontline of hospital treatment to the frontiers of scientific research, there is constant good news, which brings us hope. Samples collected in various places are sequenced to obtain viral gene sequences [1]. When the gene sequences in various places are brought together, the number of viral gene sequences becomes very large. Today there are many tools or online tools [2] to help us analyze the similarity or other characteristics of gene sequences, but when we want to quickly learn some basic information such as the distribution of a gene in all sequences, it becomes very difficult, and it often takes some time. There are 15 modules in the metagenomic analysis system MAS, which can provide unique functions to support a wider range of applications. This paper is mainly related to the A3 and C1 modules.

Aim of the Study

This study focuses on the difficulty of quickly obtaining the specified features in a large number of sequences, and hopes to visualize the results for easy understanding. First, variant measurement [9] is used to project viral gene sequences onto 1D histograms, and find suitable parameters through continuous exploration. This powerful mapping mechanism can be further explored to resolve any types of big data collections for categories and content-based indexing to provide supersymmetric properties to manage giant data collection over the world.

In this paper, a specific method is introduced. Multiple probability measures are mapped as relevant 1D histograms.

Citation: Tan Yao and Jeffrey Zheng. "Visualizations of Multiple Probability Measures for SARS-CoV-2 Genomes". EC Neurology SI.02 (2021): 15-18.

Materials and Methods

The materials used was this study are the SARS-CoV-2 gene sequences published in the GISAID (Global Influenza Data Sharing Initiative) database [5]. The length of each gene sequence used is greater than 29000. The figures shown in this paper were obtained using the gene sequences in table 1.

Sample	Mark	No.	Collected Date	Locality
SARS-CoV-2	a	EPL_ISL_412978	2020-01-17	China: Wuhan
	b	EPL_ISL_424356	2020-01-25	China: Beijing
	c	EPL_ISL_413455	2020-02-28	USA: Washington
	d	EPL_ISL_418990	2020-02-03	China: Hangzhou
	e	EPL_ISL _415709	2020-01-25	China: Hangzhou
	f	EPL_ISL_422491	2020-03-19	USA: New York
	g	EPL_ISL_416320	2020-01-28	China: Shanghai
	h	EPL_ISL_416323	2020-01-29	China: Shanghai

Table 1: *Samples.*

The main method is divided into five steps: preprocessing [8], segmentation, statistics, normalization and projection. The first step is preprocessing, which refers to the preparation of the downloaded viral gene sequences before entering the program, mainly to check whether there are abnormal characteristics in the viral gene sequences, and, if so, to deal with them accordingly. The second step issegmentation, which refers to dividing a single viral gene sequence into short sequences of fixed length m. In the variant maps shown in this paper, m is 32. The third step is statistics, which means counting the number of specific characters in the short sequences obtained in the second step. In the maps shown in this paper, the statistics include not only the individual numbers of A, C, G, T but also the sum of the numbers of C and G in each short sequence. The fourth step is normalization [3,4], which refers to normalizing multiple values obtained in a viral gene sequence. The last step is projection. The graph used in this paper is the 1D histogram. From the 1D histogram, you can see the frequency of each group and the relationship between the frequencies of each group.

Results and Discussion

This chapter will show and analyze the 1D histograms generated by the materials and methods introduced in the previous chapter. Figure 1 shows all the histograms, from which you can observe their basic characteristics, such as they are approximately normal distribution. These histograms have some common features, such as high in the middle, gradually decreasing on both sides and approximately symmetrical. Most of the small frequencies in these histograms have two or more groups. We can analyze the similarities and differences between these histograms to determine the distribution characteristics of a specific gene or gene combination in these sequences.

It can be observed from figure 1 that all histograms of sequence g are similar to the corresponding histograms of sequence h. Observe that the histograms of sequence g and sequence f, ga and fa are similar, gc and fc are similar, and only two corresponding histograms are similar. Therefore, it is inferred that sequence g is similar to sequences f and h, but sequence g has the highest similarity to sequence h.

Figure 1: *All projections of eight sequences.*

Some histograms with better symmetry are selected from figure 1 and shown in figure 2. The highest frequency number in the histogram is not only one set, so a histogram b_A with two highest frequency numbers is selected for display. The other three histograms in figure 2 have only one set of highest frequencies. It can be seen that they have good symmetry, with the highest group or two groups as the center, and gradually decrease toward both sides.

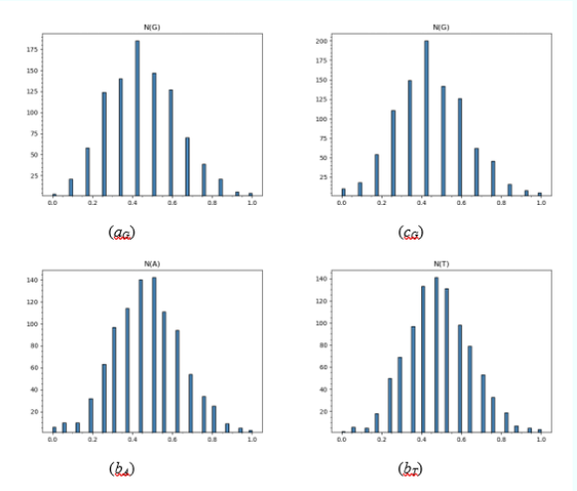

Figure 2: *Some histograms with symmetry.*

The two histograms with the center of gravity left and the center of gravity right selected from figure 1 are shown in figure 3. The center of gravity of c_A and e_C is to the left, and the gap between each group is small. c_{CG} and d_{CG} are just the opposite. Their center of gravity is to the right, and the gap between each group is slightly larger. Therefore, it is inferred that the number of A in a small segment in sequence c is abnormally large, and the number of C and G in a small segment is abnormally small. The number of a certain small segment C in sequence e is abnormally large. The number of C and G in sequence d is abnormally small.

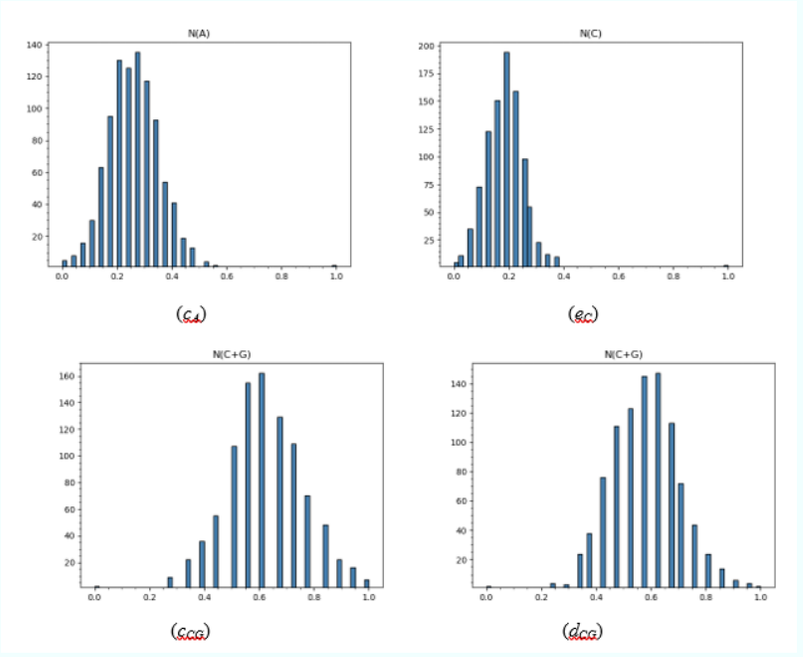

Figure 3: *Some histograms with center deviation.*

Conclusion

Multiple probability [7] measures are mapped as relevant 1D histograms. The frequency of each group and the relationship between the frequencies of each group can be observed. By projecting a large number of SARS-CoV-2 genomes, it is possible to analyze the regularity and anomaly of the number of certain symbols or symbol combinations in the sequences. The histograms obtained from most sequences projections have similarities, and they exhibit an approximately normal distribution. Variant map reflecting more diverse information will continue to be explored in subsequent studies.

Acknowledgements

The authors would like to thank NCBI, GISAID, CNGBdb, Nextstrain and Zhigang Zhang for provide 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 (2018ZI002), NSF of China (61362014), Yunnan Advanced Overseas Scholar Project.

Bibliography

1. Jongsik Chun., *et al.* "Eztaxon: a web-based tool for the identification of prokaryotes based on 16s ribosomal rna gene sequences". *International Journal of Systematic and Evolutionary Microbiology* 57.10 (2007): 22592261.
2. James R Cole., *et al.* "The ribosomal database project (rdp-ii): sequences and tools for high-throughput rrna analysis". *Nucleic Acids Research* 33.1 (2005): D294D296.
3. Karl J Friston., *et al.* "Spatial registration and normalization of images". *Human Brain Mapping* 3.3 (1995): 165189.
4. Anil Jain., *et al.* "Score normalization in multimodal biometric systems". *Pattern Recognition* 38.12 (2005): 22702285.
5. Yuelong Shu and John McCauley. "Gisaid: Global initiative on sharing all influenza data from vision to reality". *Eurosurveillance* 22.13 (2017).
6. Wenling Wang., *et al.* "Detection of sars-cov-2 in different types of clinical specimens". *The Journal of the American Medical Association* (2020).
7. Jeffrey Zheng. "Variant Construction from Theoretical Foundation to Applications". *Springer Nature* (2019).
8. Jeffrey Zheng and Chris Zheng. "Biometrics and knowledge management information systems". *In Variant Construction from Theoretical Foundation to Applications* (2019).
9. Jeffrey Zhi J., *et al.* "A framework to express variant and invariant functional spaces for binary logic". *Frontiers of Electrical and Electronic Engineering in China* 5.2 (2010): 163172.