Functional Group Decomposition of Multiple Coronaviruses on Variant Maps

Liuyun Du, Jeffrey Zheng

Abstract Different coronaviruses can be identified as three categories: common coronaviruses, fatal coronaviruses, and domestic coronaviruses. It is convenient to generate various visual results for their RNA sequences on variant maps. In this paper, a functional group measurement method is proposed to combine discrete mathematics and computational technologies on the A2 module of the MAS. Various samples are processed by this scheme and interesting results can be observed. The projections of the segmented groups on each coronavirus compared with the projective effects on different coronaviruses in 2D maps of coordinate systems are shown by statistical measures on the density matrix with similarity and dissimilarity properties for further exploration.

Keyword RNA sequence, coronavirus, functional group, variant model, visualization

Liuyun Du
Software Engineering
Key Laboratory of Quantum Information of Yunnan
Key Laboratory of Software Engineering of Yunnan
Yunnan University, Kunming e-mail: 12017002070@mail.ynu.edu.cn

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

Research Background

Coronavirus is a kind of RNA virus with envelope and linear single positive strand genomes. It accounts for approximately 1/100000 of the human genome data and approximately 1/150000 of the E.coli genome data. The coronavirus genome data are as small as 26KB and as large as 30 KB. There is no large difference in the data length, but it is the largest single stranded RNA virus among all [1]. Biochemical performance is mainly able to cause human and many livestock and animal diseases.

At present, there are seven kinds of coronaviruses: HCoV-NL63,HCoV-HKU1, HCoV-OC43,and HCoV-229E, which can cause different degrees of human infection, but the mortality rate is almost zero [2]. The other three kinds are fatal coronavirus SARS-CoV, MERS-CoV and SARS-CoV-2. All of the above seven species have one thing in common: they are all human coronaviruses, and there is also a domestic coronavirus, such as porcine deltacoronavirus (PDCoV). The experimental data in this paper were also analyzed and processed. The mortality of this disease among domestic animals and pigs is very high, 50%-100% [3].

Currently, the mainstream research on coronavirus is not to define them by various morphological characteristics and biochemical reactions but to analyze and process coronavirus by combining bioinformatics. One of the important fields of bioinformatics is to use mathematical methods to model and combine them with powerful tools in computer science to explore the mysteries of biology. The two basic processes that have to be involved are the acquisition of relevant data and the analysis behind it. These two processes are equally important. The acquisition of data is mainly through molecular biotechnology and then through high-throughput and other similar sequencing means to explore the gene sequence data of the essence of life.

The analysis of data uses the high-efficiency processing technology of computers to analyze and process the biological data to obtain their essential characteristic information and observes many results of data analysis to find the law of life. Of course, with the progress of sequencing technology, most algorithms have been unable to catch up with the data explosion speed. This requires researchers to think more closely and consider the desire to achieve low time complexity in an all-round way. However, its job is to extract effective information from nucleotide and protein sequences. This task can be transformed into the study of sequence essence, starting from four bases, to explore the mystery.

In this paper, based on geometric statistics, we present two-dimensional images of coronaviruses and analyze their characteristics. This method belongs to the A2 function module of the MAS and describes a kind of color RNA sequence visualization model. This model is based on the variant construction [9]. The variant construction is a new system composed of variant logic, measurement and visualization models, which can be used to analyze gene sequences under the condition of variations. it has been used in the random detection of sequence cipher, and also used in the medical field to detect different diseases visualization of human ECG data can not only solve the problem of information loss and degradation, but also realize dense visualization [4,5]. It needs less plane space to visualize DNA se-

quences with larger data. It mainly uses the relationship between A, T, G and C to process the whole DNA sequence and finally forms 2D maps in space. Through the processing of this module, we can see the different projection effects of different coronaviruses in the coordinate system, and according to the decomposition of coronavirus gene sequence, observe its functional gene segmentation, contact the statistical characteristics of bases, and obtain the internal rules.

Data and Methods

Data Sources

All coronavirus data in this paper are collected from the NCBI (https://www.ncbi.nlm.nih.gov/) and GISAID (https://platform.gisaid.org/), and a series of corresponding serial login numbers are shown in the table below.

category	code name	virus serial number
Common coronavirus	HCov-229E	NC_002645
	HCoV-NL63	NC_005831
	HCoV-OC43	NC_006213
	HCoV-HKU1	NC_006577
Fatal coronavirus	MERS-CoV	JX869059
	SARS	AY508724，AY485277
	COVID-19	EPIa-SL-424344
Domestic animal coronavirus	Porcine delta coronavirus	KX022605

Main Methods

In the gene sequence, base pairing follows a strict complementary symmetry. From the 15 parameters of the probability measure $(A,T,G,C,A+T,A+G,A+C,T+G,T+C,G+C,A+T+G,A+T+C,A+G+C,T+G+C,A+T+G+C)$, a variety of one-dimensional or multidimensional visualization modes can be formed [8]. According to prior experiments and biological knowledge, to obtain better graphical results, this paper mainly selects the two-dimensional visualization framework, focusing on the statistics of the probability of A+T and A+G parameters, and then, in the variant model, a 2D map is generated according to the count of the coresponding numbers [7].

Architecture

In this paper, we first obtain the relevant data from the NCBI website as the input and then filter the data to obtain the required format. Finally, we process all sequences with the same segmentation and method, and each data point can obtain a 2D map output. The whole frame structure of the RNA gene sequence is shown in Figure 1.

As the basic experimental data, the RNA gene sequence is divided into several equal length subsequences, and the data in each subsequence are calculated according to the measurement parameters.

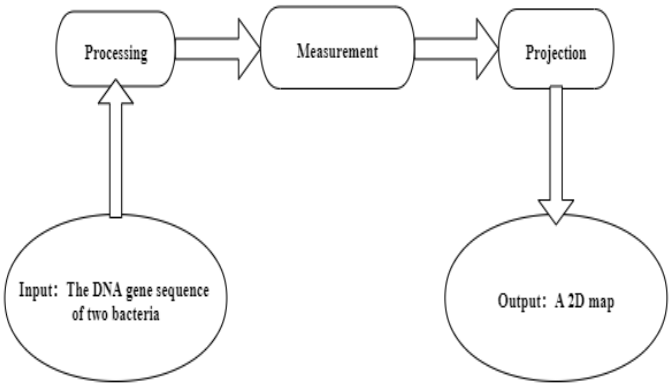


Fig. 1: Frame structure

Parameter setting:
m: Segment length of RNA sequence (m = 30)
Number$(A+T)$=Number(A)+Number(T)
Number$(A+G)$=Number(A)+Number(G)
Map to figure x, y settings:
X=Number$(A+T)$
Y=Number$(A+G)$

Read the RNA sequence, select n genes as the selected RNA sequence, and as the input data, a total of M segments. After quantitative statistics, the normalized measure is obtained. Through the Number $(A+T)$, Number $(A+G)$, the abscissa and ordinate x and y of each point can be obtained, and the position of this point in Cartesian coordinates can be determined. Each coordinate point is taken as the input value of the image projection part, and all selected RNA sequences are collected for image projection. Finally, the characteristic distribution map of ancient bacterial RNA sequences is obtained. The deeper the color in the two-dimensional image formed by mapping, the denser the distribution of AT and AG.

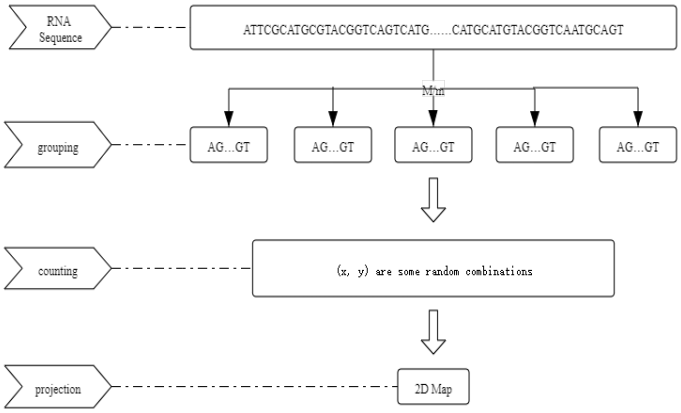


Fig. 2: Measurement module

The following example shows the projection results of the module to deepen the understanding of the subsequent visualization results. The range of the two coordinate axes and chromatographic bars is (0,30). The larger the projection value is, the closer it is to yellow, and the smaller the value is, the closer it is to blue. The points in the coordinate system represent (Number$(A+T)$, Number$(A+G)$).

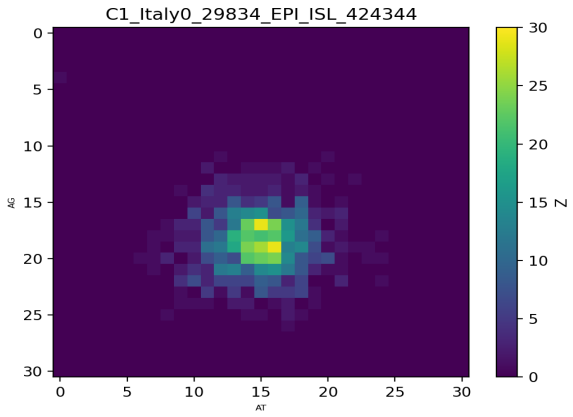
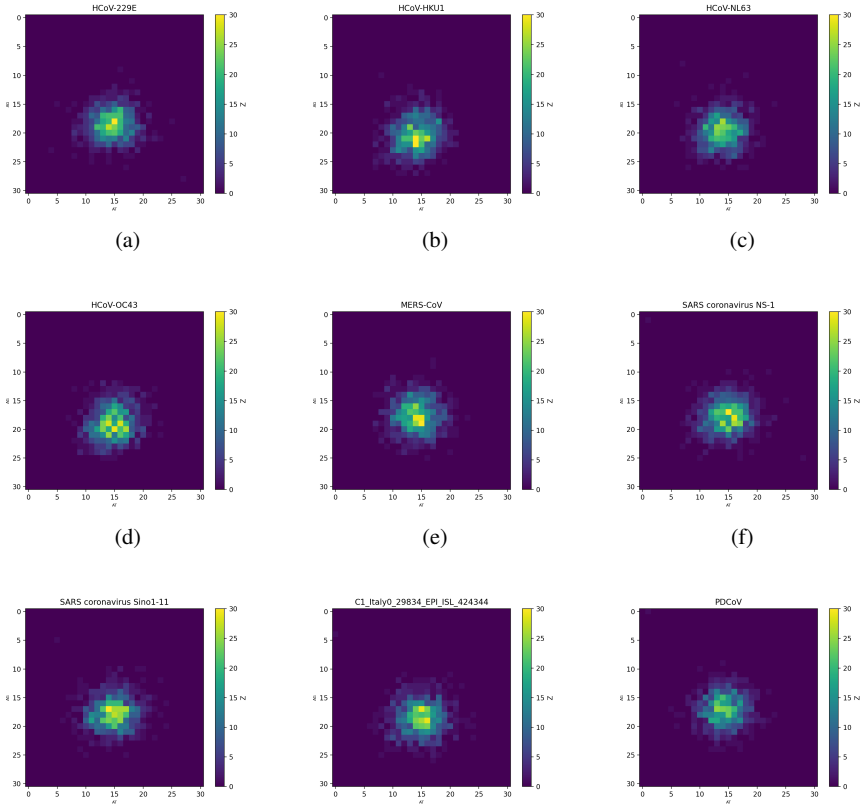


Fig. 3: A projection example

Visualization Results and Analysis



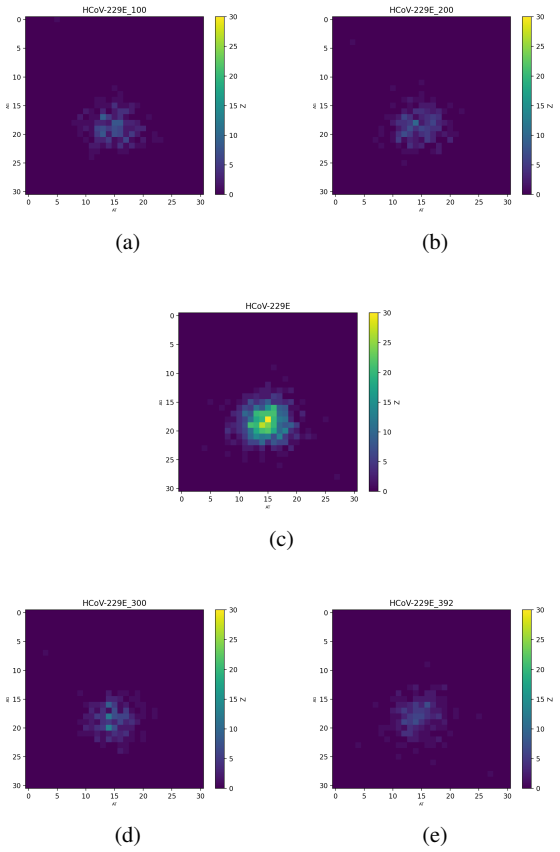
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 4: Nine different coronaviruses

Figure 4 shows nine different coronaviruses, the first four (Figure 4(a)-(d)) are common coronaviruses, the fifth to the eighth (Figure 4(e)-(h)) are fatal coronaviruses, and the last one (Figure 4(i)) is domestic coronaviruses. It can be seen from the above pictures that, in general, the distribution areas of all coronaviruses in the coordinate system are relatively similar, and the higher the mortality rate of coronaviruses, the higher the projection area.



(a) (b) (c) (d) (e)

Fig. 5: HCoV-229E segmented projection and total projection

Figures 5 and 6 show the functional genes of the HCoV-229E segment, which is convenient to observe their distribution characteristics. 100 lines of data are used as segments, and it can be seen that the coverage rate of the 100-200 segment (Figure 5(b)) of HCoV-229E is the closest to that of the whole gene sequence, as well as that of HCoV-NL63 in Figure 6.

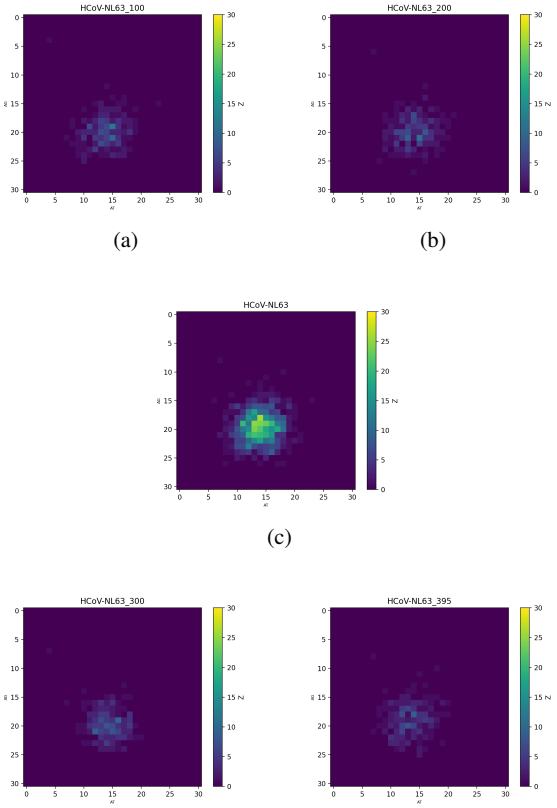


Fig. 6: HCoV-NL63 segmented projection and total projection

Result Discussions

Because the influence of coronavirus on humans cannot be ignored, it plays an important role in the field of bioinformatics. In this paper, the study of coronavirus is to analyze the expression of coronavirus diversity and explore the existence of functional genes by using the probability statistical variable value model. This method can describe the distribution characteristics of different kinds of coronaviruses quickly, conveniently, concisely and intuitively.

Conflict Interest

No conflict of interest has been claimed.

References

1. Duffy.S, Shackelton.L,Holmes.E *Rates of evolutionary change in viruses: patterns and determinants*, Nat Rev Genet,2008(9):267276.
2. Zhao Yanjie, Tan Wenjie *Genome structural characteristics and molecular detection of the Middle East respiratory syndrome coronavirus*, Chinese Journal of preventive medicine, 2015(5):461-464.
3. Chen yunhuan,Wu Yan'an*Research progress of two new human coronaviruses HCoV-NL63 and HCoV-HKU1*, Medical Review, 2010-16(17):2596-2598.
4. Ji Yan*Research on Visualization Application of ECG data series based on variant Measurement*, Yunnan University, 2016.
5. Leng Lihua, Jeffrey Z.J. Zheng*Visualization of ECG sequence of sinus arrhythmia*, Computer science,2016,43(S2):183-185.
6. Wan Zhu, Jeffrey Z.J. Zheng*Visualization of one-dimensional segmented measurement and distribution of DNA sequences*, Journal of Yunnan University (NATURAL SCIENCE EDITION),2013,35(S2):1-6.
7. Mao Yuyuan,Jeffrey Z.J. Zheng,Liu Wenjia *Mapping Whole DNA Sequence on Variant Maps*, 1037-1040. 10.1145/3110025.3110140, 2017.
8. Jeffrey Z.J. Zheng, Christian H.H. Zheng *Variable measurements and visualized statistical distributions*, Acta PHOTONICA Sinica, 2011,040(009):1397-1404.
9. Jeffrey Z.J. Zheng *Variant Construction from Theoretical Foundation to Applications*, 10.1007/978-981-13-2282-2, 2018/9/12