3D Visualizations of Multiple Coronaviruses on Whole Genomes

Zhongwei Zhang, Tingyan Duan, Jeffrey Zheng

Abstract COVID-19 triggered by SARS-CoV-2 has become a common problem faced by people all over the world. With the development of bioinformatics and the breakthrough progress of gene technologies. It is a challenging topic to use genomic datasets for SARS-CoV-2 research. In this paper, a 3D visualization method is proposed to show the A9 module of the metagenomic analysis system MAS. Seven coronaviruses of genera were illustrated and briefly analyzed. Comparing the visualization results, various SARS-CoV-2 genomes were represented as 2D and 3D maps under different conditions. Through related specific projections, the characteristics of the coronavirus can be observed intuitively from the projection results to provide an effective viewpoint for studying viral genomes.
Keywords: coronavirus, genome, 3D map, 2D map, visualization, projection

Zhongwei Zhang
School of Software, Yunnan University e-mail: 1690619933@qq.com

Tingyan Duan
Information Engineering College, Nanyang Vocational College Of Agriculture, China e-mail: 1181268816@qq.com

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

This work was supported by the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018ZI002)

Introduction

In December 2019, a group of people with new coronary infections were discovered. The full genome sequence of the virus was obtained on January 29, 2020, with a total length of 29847 bp. On February 11, the World Health Organization named the new coronary pneumonia disease "COVID-19", and the International Committee of Viral Taxonomy named the virus "SARS-CoV-2".

As of April 28, the number of diagnoses worldwide has exceeded 3026981. SARS-CoV-2 [1] has now seriously threatened the health of the global public, and it has attracted widespread attention from people around the world. Research on SARS-CoV-2 at home and abroad is also increasing.

The outer layer of the coronavirus has an envelope, and the shape is spherical or elliptical, with polymorphism. The genome is a linear single-stranded RNA virus, which is a large class of viruses that are ubiquitous in nature [2–6].

The Coronaviruses are divided into four genera by the International Virus Classification Committee: alpha, beta, gamma and delta. Among them, HCoV-OC43, HCoV-NL63, HCoV-229E, HCoV-HKU1, SARS-CoV, MERS-CoV and SARS-CoV-2 belong to beta genera [7–10].

Studying the similarities and differences between these seven coronaviruses from the perspective of genome sequence [11–15] visualization plays a vital role in preventing and controlling new coronaviruses and preventing the spread of disease.

Data visualization technology is a technology that displays abstract data in an intuitive graph or image, thereby facilitating research and analysis [16,17]. There are many visualization methods for genomic sequences: most of the genomic sequence visualization models are implemented by DNA walking technology.

For example, the Gates-Nandy model [18] has an information degradation problem. To solve the problem of degradation and data loss, the researchers proposed a CGR model [19], a three-dimensional visualization model [20], and a worm model based on the Gates-Nandy model [21]. In 2003, Randic [22] proposed a spectral visualization model, which is different from the Gates-Nandy model. It consists of four parallel lines with the same distance, and four bases (adenine A, thymine T, bird The connection of purine G and pyrimidine C) also solves the problem of information degradation.

However, the above visualization models for genomic sequences are not suitable for processing long DNA sequence data, and the analysis methods for visualization results are not universal. In 2014, Feng Haiqing et al. proposed a grayscale image-based DNA sequence visualization model [23]. This model converts one-dimensional DNA sequence information into two-dimensional 256-color gray by encoding four bases. The degree of image greatly compresses the length of the DNA sequence visualization, and has a high spatial tightness. However, the visualization of the model has noise, which is not conducive to researchers observing more effective information.

According to the properties of the genome sequence, a new visualization method of the genome sequence is developed based on the research idea of data visualization technology and based on a variant logic system [24–28]. By adjusting the

parameters, it can adapt to the genomic sequence with a large amount of data. By processing the seven coronavirus genome sequences, new graphical results can be obtained, and the resulting graphs clearly converge. From the results, we can find the correlation between the data and clarify the changing rules and phenomena, so that we can observe some characteristics of coronavirus from the perspective of variant.

Materials and Methods

Structurally speaking, variant logic is composed of four primitives. The four primitives represent four different states. There are symmetrical states and complementary states between two pairs. All states are combined to form the entire space. Segmentation is a major feature of variants. Dividing the data into equal segments can quickly process large amounts of data. The state value of each segment can be obtained by calculation. The visual model based on the variant system is to illustrate the state set of each segment.

The visualization model processing flow is shown in Figure 1:

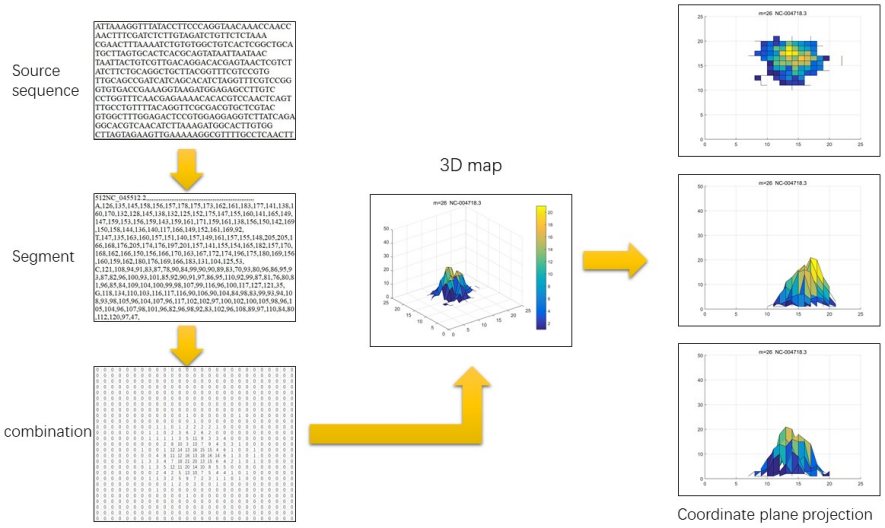

Fig. 1 Processing flow chart

The processing flow is as follows: input the viral genome sequence, and segment the genome sequence according to the segmentation value m to obtain the segmentation result. Calculate the distribution of {A, G, C, T} in the segmentation results respectively to obtain the elementary states $\{X_A, X_G, X_C, X_T\}$. After the elementary states are obtained, two combinations can be selected between the elementary

states to obtain superposition states$X_{(A+T)}, X_{(A+C)}, X_{(A+G)}, X_{(G+C)}, X_{(G+T)}, X_{(C+T)}$. Choosing the values in the primitive state and the superposition state to project can obtaint the graphical results.

Introduction of the main modules in the model:

(1) Segmentation result: The segmentation result is affected by the segmentation value. The segmentation value is usually represented by m, which represents the number of bases in each segment after the entire sequence is divided equally. The genome sequence is divided into n/msegments by changing the size of the segment value m, and each segment has m bases. Controlling the size of m can adjust the resolution and effective area of the result.

(2) Primitive state: the proportion of four bases in each segment is the primitive state, which represents the proportion of four bases in the genome sequence. The four primitive states in the variable-value system have a substitution and complementarity relationship, which perfectly fits the principle of complementary pairing between bases in the genome. The four symbols X_A, X_G, X_C, and$X_T$are used to represent the states of the four bases A, G, C, and T in each segment.

(3) Superposition state: Each primitive state can be combined with each other, and can also be combined with three or four primitive states. A state such as this is called a superposition state. Each superposition state has a different meaning. There are six superposition states combined by two elementary states, which are$X_{(A+T)}, X_{(A+C)}, X_{(A+G)}, X_{(G+C)}, X_{(G+T)}, X_{(C+T)}$ said.

(4) Graphical results: This experiment uses three-dimensional charts and two-dimensional projection charts to analyze the data.

Three-dimensional graph: It can display the features of primitive states and superimposed states. The three-dimensional map has a larger space capacity and can display more features. Two kinds of projections are randomly selected from the four primitive states and superposition states, respectively, as the x-axis and y-axis, and their values are accumulated at the corresponding positions to generate the z-axis. Finally, a three-dimensional diagram can be generated. Here select $X_{(A+T)}$ and$X_{(A+G)}$ to generate a three-dimensional map.

Two-dimensional projection map: The three-dimensional image can clearly see the overall structure of the viral genome sequence, and the specific details can be projected onto a two-dimensional plane for observation. Projecting the 3D map onto other planes of the coordinate system generates a 2D projection of the 3D map.

Projecting various genomic sequences into three-dimensional space can observe its overall features from various angles, and the details contained in the overall features can also be displayed in the two-dimensional projection diagram generated by it.

Results and Discussion

Data introduction

In the variant visualization model, better analyze the spatial distribution characteristics and cycle characteristics of the seven coronavirus genome sequences, they are expressed in two-dimensional and three-dimensional maps, respectively. To ensure the reliability of the data, the whole genome sequences of seven viruses were downloaded from NCBI. The corresponding sequences and sizes of virus names are as follows:

Table 1 Whole genome sequence information of seven viruses

Name	Serial number	Data length
HCoV-OC43	NC-006213.1	30741 bp
HCoV-NL63	NC-005831.2	27553 bp
HCoV-229E	NC-002645.1	27317 bp
HCoV-HKU1	NC-006577.2	29926 bp
SARS-CoV	NC-004718.3	29751 bp
MERS-CoV	NC-019843.3	30119 bp
SARS-CoV-2	NC-045512.2	29924 bp

From the data information, it can be seen that the lengths of the whole genome sequences of the seven coronaviruses are not much different, on an order of magnitude. The genome sequences of living organisms are all in the millions. Although the genome sequences of coronaviruses are not long, they are approximately 30,000. Using variant results to project data into an $n \times n$ matrix, useful information can be observed simply and intuitively.

Projection selection

In the process of variant processing, each link can be adjusted, and this feature makes the method adapt to various data. However, the impact of each variable on the results makes it necessary to unify the comparison. Therefore, the appropriate value is calculated by the control variable method, and then the results obtained by the fixed parameters are compared. The two main parameters that affect the result are the m value and the other is the selection and combination.

The three-dimensional space is affected by the value of m in all aspects. The most direct impact is the size of the resulting graph and the speed of processing the data. We use the mean, variance and standard deviation as indicators to determine the effect of the m value on the three-dimensional space. The table lists the mean, variance and standard deviation of the seven coronavirus sequences at $m = 26$.

The mean value reflects the amount of information contained in the unit space in the three-dimensional space, and the variance and standard deviation represent the degree of dispersion of the data points in the data set. It can be seen from the table that the indicators of the seven coronaviruses are not much different. Therefore, it has better observation effect when $m = 26$.

By comparing the display area of the results, select the combination of $X_{(A+T)}$ and $X_{(A+G)}$ to obtain a three-dimensional graphical result. Therefore, for the three-dimensional graphical result, select the result of the combination of $X_{(A+T)}$ and $X_{(A+G)}$ when $m = 26$.

Table 2 Mean, Variance and Standard deviation of seven coronaviruses

Name	HCoV-OC43	HCoV-NL63	HCoV-229E	HCoV-HKU1	SARS-CoV	MERS-CoV	SARS-CoV-2
Mean	1.62	1.45	1.44	1.58	1.57	1.58	1.57
Variance	26.64	23.61	23.61	24.25	25.45	26.01	25.97
Standard deviation	5.16	4.85	4.85	4.92	5.04	5.10	5.09

Results Show

Under the fixed parameters $m = 26$, $X_(A+T)$ and $X_(A+G)$ combination, seven kinds of coronavirus obtain a three-dimensional map: Figure 2

Rotate and project the 3D image to obtain the projection results of the 3D image on other planes. The specific value of the three-dimensional projection can be seen by observing the image of the plane projection.

Figure 3 is the result of projecting a three-dimensional image onto the xy plane.

Figure 4 is the result of projecting a three-dimensional image onto the xz plane,

Figure 5 is the result of projecting a three-dimensional image onto the yz plane.

Result analysis

As seen from the three-dimensional graphical results, the graphical results of the seven coronaviruses are clustered in the middle and distributed around. They are all distributed in the upper left corner of the space. In particular, the peak value of e is obviously lower than that of the other six coronaviruses. Acf has a clear single peak.

The result of mapping the three-dimensional graphic results to the xy coordinate plane. From the figure, it is more accurate to observe the single peak position of acf. The distribution of the double peaks of several other coronaviruses is also different.

The result of mapping the three-dimensional graphic results to the xz coordinate surface and the yz coordinate result supplements the deficiencies of other surfaces.

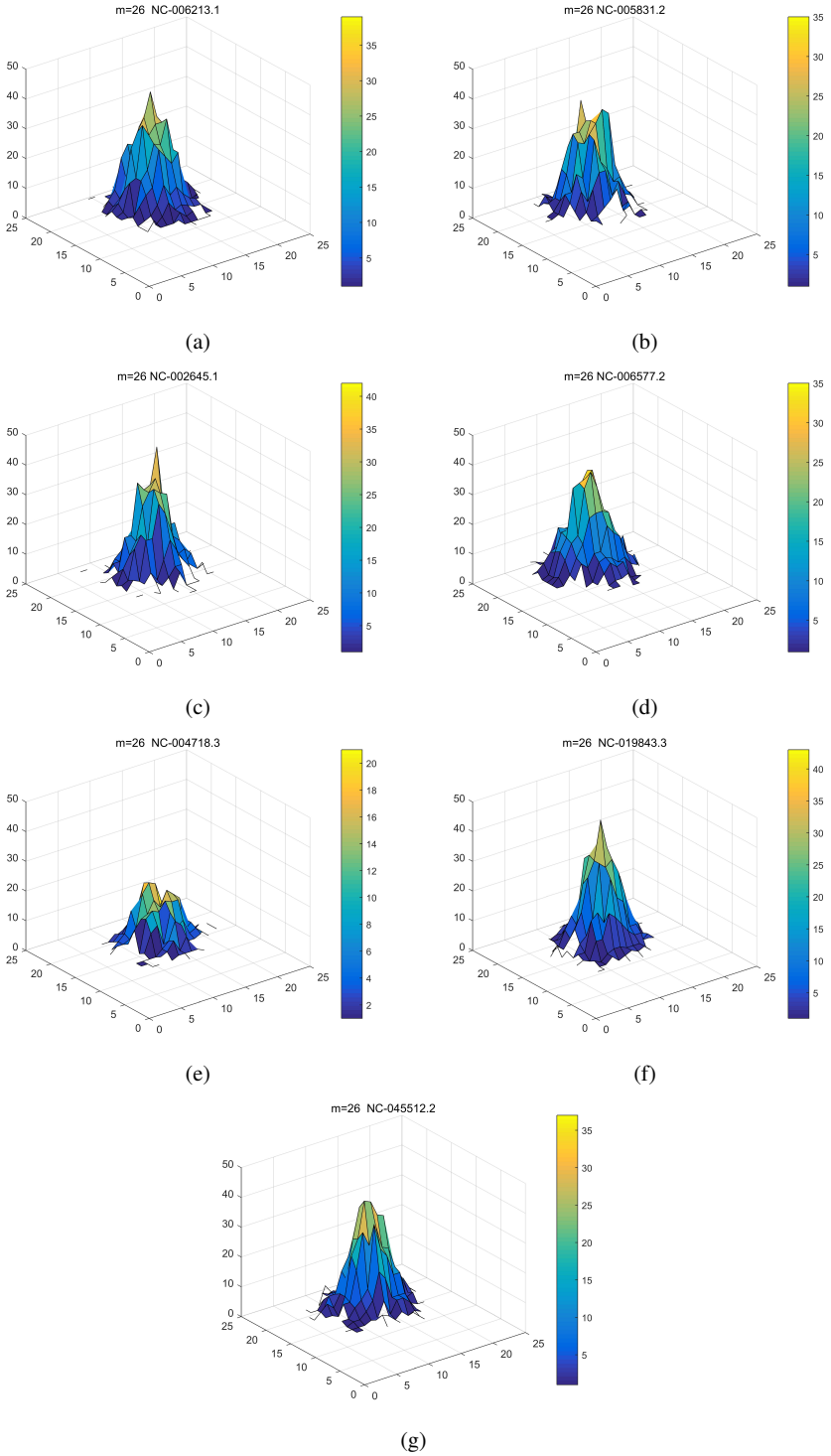

Fig. 2 m=26,3D graphical results of HCoV-OC43, HCoV-NL63, HCoV-229E, HCoV-HKU1, SARS-CoV and MERS-CoV

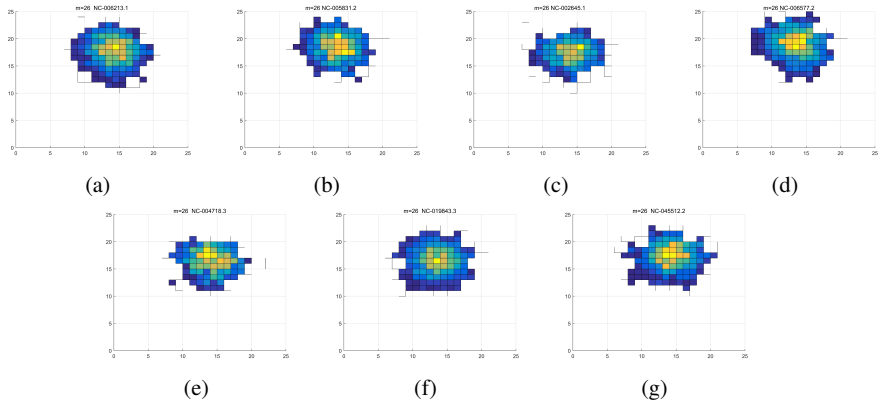

Fig. 3 m = 26,HCoV-OC43, HCoV-NL63, HCoV-229E, HCoV-HKU1, SARS-CoV and MERS-CoV three-dimensional images are projected onto the xy plane

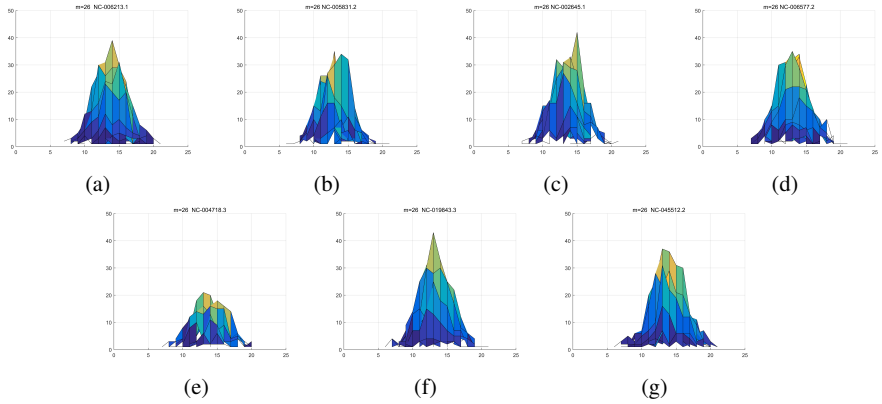

Fig. 4 m = 26, HCoV-OC43, HCoV-NL63, HCoV-229E, HCoV-HKU1, SARS-CoV and MERS-CoV three-dimensional images are projected onto the xz plane

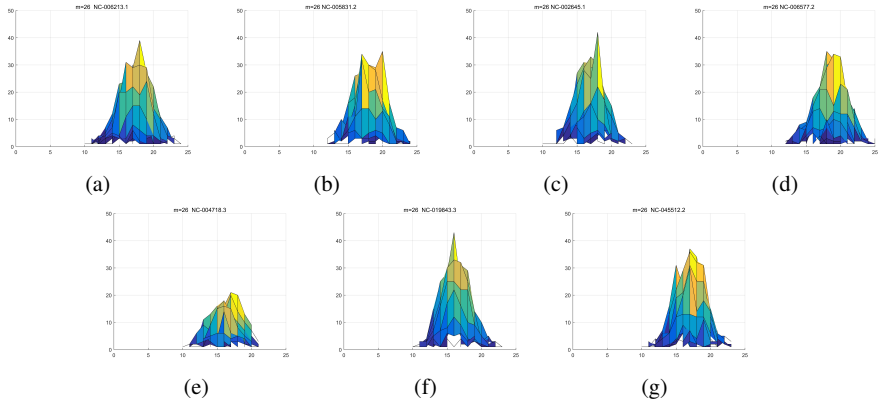

Fig. 5 m = 26, HCoV-OC43, HCoV-NL63, HCoV-229E, HCoV-HKU1, SARS-CoV and MERS-CoV three-dimensional images are projected onto the yz plane

It is precisely seen that the result value of e is 21, and the result value of c is 42 is twice that of e.

Conclusion

A variant visualization model was established for seven kinds of coronaviruses, and the data were converted into easy-to-understand graphs. The two-dimensional and three-dimensional graphs were suitable for different variant calculation stages. Calculate the m value to adjust the best visual effect. Use three-dimensional projection to observe the distribution of superimposed states. The three-dimensional image is proposed to two-dimensional to accurately observe the state of the superimposed state at various angles in the three-dimensional image. The results show that the results obtained by the visualization method based on the variant system can clearly show the connection and difference between the seven viruses. The study of new coronaviruses provides a new non-biological method.

Conflict Interest

No conflict of interest has been claimed.

Acknowledgements

We thanks to the NCBI platform and the researchers for providing the genomic data.

References

1. Yuen KS,Ye ZW,Fung SY,Chan CP,Jin DY. “SARS-CoV-2 and COVID-19:Themostimportant research questions”. Cell Biosci 3.16(2020).
2. Yijing Yang and Yunwen Hu. Molecular epidemiological study of human coronavirus OC43 in Shanghai from 2009-2016. Chinese Journal of Preventive Medicine 52.1(2018):55-61.
3. Zhong-tian QI. Severe acute respiratory syndrome coronavirus 2 and coronavirus disease 2019. Academic Journal of Second Military Medical University 41.2(2020):117-121.
4. Ming YIN.,et al, Epidemiological patterns of coronavirus diseases. Chinese Journal of Multiple Organ Diseases in the Elderly 19.3(2020):187-190.
5. Jie Yan.,et al, 2019 novel coronavirus (2019-nCoV) and 2019-nCoV pneumonia. Chinese Journal of Microbiology and Immunology 40.1(2020):1-6.
6. Kerui Weian. Yue Gong. Zhixiang Shi.,et al, Current Status of Research on Coronavirus. China Biotechnology 40.1(2020):1-20.
7. Yue Gong.,et al, A Bibliometric Analysis on Coronaviruses.China Biotechnology 40.1(2020):21-37.

8. Dan-yi SHI.,et al, Genetic analysis on the homology between severe acute respiratory syndrome coronavirus 2(SARS-CoV-2) and common domestic animal coronaviruses. Jiangsu Journal of Agricultural Sciences 36.1(2020):251-253.
9. chengping Lu. Comparison of SARS Virus and Animal Coronaviruses. VIROLOGICA SINICA 18.3(2003):307-309.
10. Ruo-ying WANG and Ya-li WANG. Epidemiological characteristics of Middle East respiratory syndrome, 2015. Practical Preventive Medicine 24.2(2017):193-195.
11. Yaya Hu.,et al, Research progress on the reduced-representation genome sequencing technique. Journal of Jiangsu Normal University(Natural Science Edition) 36.4(2018):63-68.
12. Bin LIU.,et al, Genomic structure and protein profiles of 2019 novel coronavirus. Journal of Microbes and Infections 15.1(2020):52-57.
13. Wanqi ZHANG.,et al, Sequence and Structural Analysis of the Complete Genome of the Bovine Coronavirus BCV-Aks-01 Strain in Southern Xinjiang, China.Chinese Journal of Virology 33.4(2017):583-592.
14. guangxing LI and Long PAN. Research development of genome structure and related protein of coronavirus. Journal of Northeast Agricultural University 44.9(2013):149-154.
15. huili Liu and chengping Lu. The Genome and Encoded Proteins of Coronavirus. PROGRESS IN VETERINARY MEDICINE 25.2(2004):1-3.
16. Shi-zhen LIAN.,et al, Research on the application of visualization technology in genome sequencing. Journal of Hunan City University(Natural Science) 25.1(2016):133-134.
17. Kan Liu.,et al.” Data Visualization Research and Development”. COMPUTER ENGINEER 28.8(2002):1-2+63.
18. NANDY A. A new graphical representation and analysis of DNA sequence structure. I. Methodology and application to glob in genes. Current Science 66.4(1994):309-314.
19. MILAN R. Another look at the chaos-game representation of DNA. Chemical Physics Letters 456(2008):84-88.
20. NAFISEH J and IRANMANESH A. C-curve: A novel 3D graphical representation of DNA sequences based on codons. Mathematical Biosciences 241.2(2013):217-224.
21. MILAN R. Graphical representation of DNA as 2-D map. Chemical Physics Letters 386(2004):468-471.
22. Randic M, Vracko M, Lers N, et al. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters 368.1-2(2003):1-6.
23. Haiqing FENG and Zuhong LU. A novel visual modal of DNA sequence based on image. China Journal of Bioinformatics 12.2(2014):133-139.
24. Zheng J., Zhang W.,et al, (2019) Variant Map System to Simulate Complex Properties of DNA Interactions Using Binary Sequences. In: Zheng J. (eds) Variant Construction from Theoretical Foundation to Applications”. Springer, Singapore (2019): 353-377.DOI:doi.org/10.1007/978-981-13-2282-2-23.
25. Zheng J.(2019)Variant Logic Construction under Permutation and Complementary Operations on Binary Logic .In: Zheng, J., Ed., Variant Construction from Theoretical Foundation to Applications”. Springer, Singapore (2019): 3-21.DOI:doi.org/10.1007/978-981-13-2282-2-1.
26. Jeffrey Zheng, Variant Construction from Theoretical Foundation to Applications. Springer Nature 2019.https://www.springer.com/in/book/9789811322815.
27. Jeffrey Zheng, Variant Construction Theory and Applications. Theoretical Foundation and Applications”, Science Press 2020 (Chinese).
28. Jeffrey Zheng and Chris Zheng, Biometrics and Knowledge Management Information Systems, Chapter 11: Variant Construction from Theoretical Foundation to Applications, Springer Nature (2019):193-202.