Protein Coding of Variations on SARS-CoV-2 Genomes in Various Regions

Tao Li, Jeffrey Zheng

Abstract In this paper, COVID-19 cases in different regions are used for comparison. The related genomes of SARS-CoV-2 are segmented and replaced with sequence operations under protein coding scheme on the A3 module of the MAS. Using protein coding schemes, genomes are transformed and projected as measuring sequences as a vector that can be visualized in maps from two different perspectives: the elements of the gene sequence and the position of the element sequence, so as to interpret the genome more comprehensively. Through a series of linear diagrams, it is convenient to compare and analyze the genomes of the samples collected in different regions more intuitively, which may be conducive to further data mining of genomic information and refined explorations of COVID-19 for patients.
Keyword COVID-19, variant construction, protein coding, replace, sequence operation, visualization

Tao Li
Yunnan University, Kunming, e-mail: 1977675165@qq.com

Jeffrey Zheng
Yunnan University e-mail: conjugatelogic@yahoo.com

Introduction

In December 2019, pneumonia caused by COVID-19 broke out in Wuhan, China. Its clinical symptoms are different from the SARS outbreak in 2003, so it is inferred that the virus may be a new variant of coronavirus [1]. Today, in just four months, new coronaviruses have swept the world. The number of countries or regions affected has already exceeded one hundred. The cumulative number of confirmed cases has exceeded two million, and the number of deaths has exceeded two hundred thousand.

Interpretation of the viral gene sequence is the key to defeating the epidemic. It can provide us with methods and ideas to fight the epidemic and provide a corresponding basis for its treatment.

Aim of the Study

Using the variable value theory system framework, the element statistics and element position statistics of COVID-19 gene sequences can be displayed through visualization methods to observe and analyze the distribution of data features.

Materials and Methods

In this paper, the gene sequence is used as the input, and the corresponding value map is used as the output. The architecture used in this article is mainly divided into the following modules: segmented operation module, replacement operation module, statistics module, and visualization module.
Under this architecture, the data processing flow is as shown in Fig. 1.

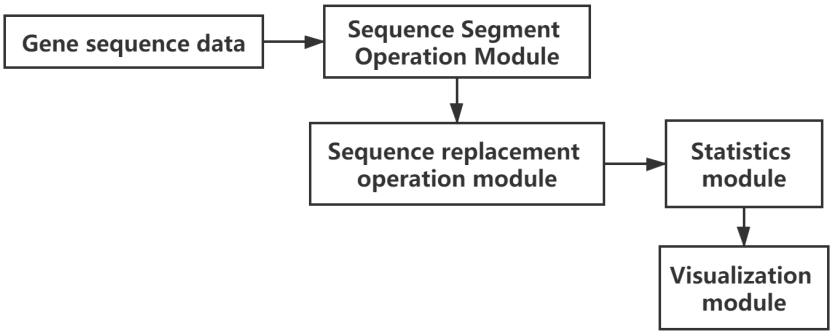


Fig. 1 Flow Chart Of Data processing

Sequence Segment Operation Module

This module segments the entire sequence of gene sequence data according to the segment length (*dlen*) into a single sequence, sequence data set with a fixed length value, where a single sequence length is less than the segment length will be discarded.

For example, the sequence data string is:
GCTTGTCAACTGCATCACATCCACGTTTACGTACACTAC

The total length of the sequence is 39; if the segment length (*dlen*) is fixed to 9, then you can obtain a sequence data set of 4 segments:

(1) *GCTTGTCAA*; (2) *CTGCATCAC*; (3) *ATCCACGTT*; (4) *TACGTACAC*.

The selection of the segment length value will be determined according to the actual length of the gene data sequence and the visualization effect.

Sequence Replacement Operation Module

The operation module mainly includes base replacement and position replacement, wherein the position replacement operation is based on the base replacement operation. According to the different properties of the four bases, any DNA sequence can be uniquely described as the distribution of three independent purines (R) and pyrimidines (Y), the distribution of amino groups (M) and carboxyl groups (K), and strong hydrogen bonds (S) and the weak hydrogen bond (W) distribution. In this paper, the three distributions are used as replacement relationships, as follows:
(1) RY: purines (R)= {A, G}; pyrimidines (Y)= {C, T};
(2) MK: amino groups (M)= {A, C}; carboxyl groups (K)= {G, T};
(3) SW: strong hydrogen bonds (S)= {G, C}; weak hydrogen bond (W)= {A, T}.

According to the corresponding replacement relationship above, one gene sequence can be mapped into three different sequences. This is the base replacement operation of the sequence, which mainly targets the elements in the sequence.
For example: the sequence $GTCCACTGGCATGGT$ can be replaced with three independent sequences:
(1) $RYYYRYYRRYRYRRY$; (2) $KKMMMMKKKMMKKKK$; (3) $SWSSWSWSSSWWSSW$.

There are four positions in any sequence after base substitution operation, as follows:
(1) RY: {RR, RY, YR, YY}
(2) MK: {KK, KM, MK, MM}
(3) SW: {SS, SW, WS, WW}

The position replacement in the sequence operation is to select a specific position relationship string as the judgment basis for replacement in all position relationships and replace the substrings equal to it in the entire data sequence with '1', otherwise replace Is '0', and finally replace the entire sequence with a sequence containing only '0' and '1'.

In summary, any data sequence that has undergone base replacement can be replaced with 4 '01' sequences in total according to different position replacement relationships, which can be used as input to the statistical calculation module.
For example, if the sequence after base substitution: $SWSSWSWSSSWWSSW$, the position replacement result is:
(1) 'SW ': 10010100010001
(2) 'SS': 00100001100010
(3) 'WS ': 01001010000100
(4) 'WW':00000000001000

Statistics Module

The statistics module mainly includes two aspects:
(1) Statistics of the number of gene sequence elements;
(2) Statistics of the position of gene sequence elements.

In the gene sequence element number statistics module, only the total number of single elements after the base replacement operation is counted, and then the measure is calculated, that is, the total number of statistics of the element is divided by the length of the sequence.

Similar to the statistical operation of the number of gene sequence elements, the gene sequence element position statistics module counts the total number of '1' characters in the '01' sequence after the position replacement operation, and then calculates its measure, divided by the total number of statistics of '1' Take the length of the sequence.

The same sequence will have different sets of statistical values due to different sequence replacement relationships.

Visualization Module

This module is the average linear visualization operation module. The sequence data set statistics are the input of the module, and a series of linear diagrams are the output of the module. The specific operation process is shown in Fig. 2.

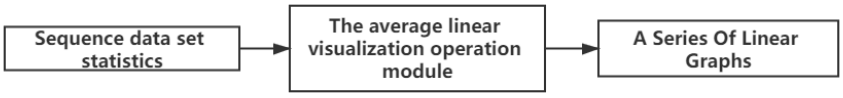


Fig. 2 Flow chart of visual processing

The average linear visualization operation uses the average value to reflect the overall situation of the data set, visualizes the distribution of gene sequence elements and the statistical distribution of element position relationships under different segment lengths, and more directly shows the distribution of gene sequences.

Data Introduction

There are 18 regions in the data involved in the article, which are divided into four independent comparison combinations, as follows:
(1) China, Australia, Brazil, Colombia, France;
(2) China, United States, Malaysia, Spain, South Africa;
(3) China, Turkey, South Korea, Peru, Sweden, Nepal;
(4) China, Greece, Iran, Israel, Italy.

All data come from the NCBI website of the American Biogene Database. At the same time, the data used are samples of data sequences submitted earlier in the region. The gene sequence ID number is as follows.

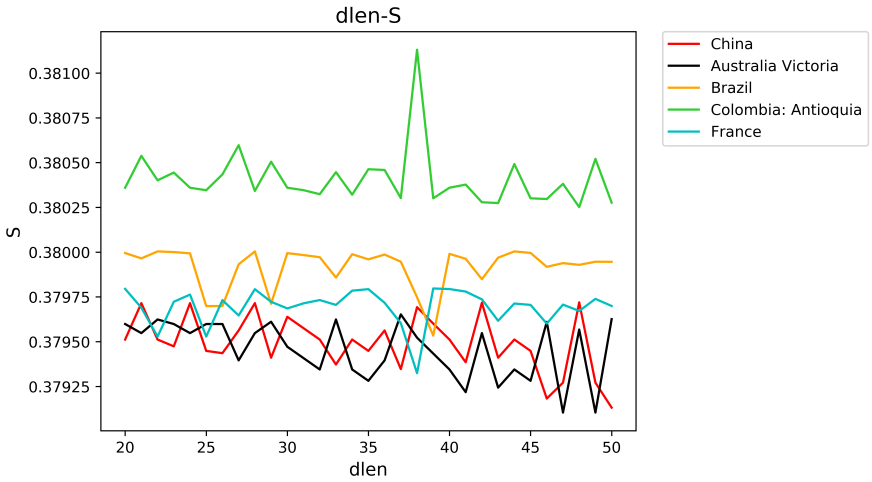
Table 1 Data Introduction

Region	Gene sequence	Region	Gene sequence	Region	Gene sequence
Brazil	*MT*126808	Australia	*MT*007544	China	*MN*908947
Colombia	*MT*256924	France	*MT*320538	Greece	*MT*328032
Iran	*MT*320891	Israel	*MT*276597	Italy	*MT*066156
USA	*MT*326173	Malaysia	*MT*372480	Spain	*MT*292570
South Africa	*MT*324062	Turkey	*MT*327745	South Korea	*MT*039890
Peru	*MT*263074	Sweden	*MT*093571	Nepal	*MT*072688

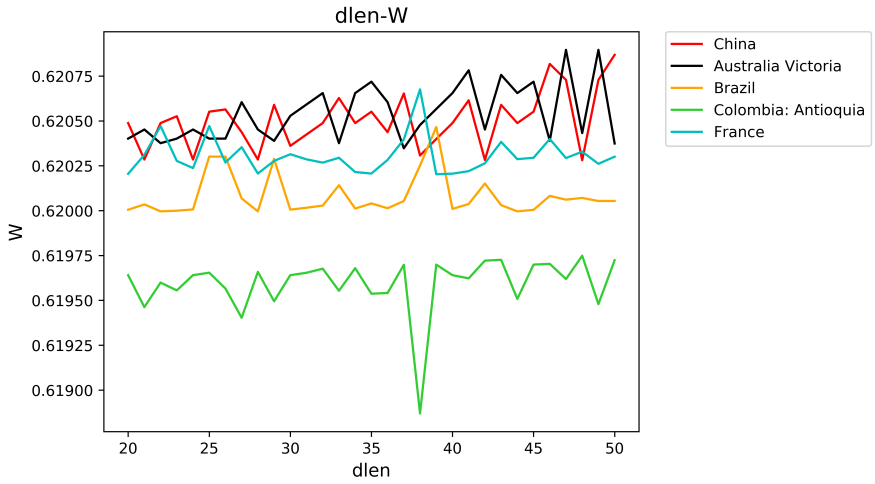
Results and Discussion

Results Display

The results of this article are mainly the content in the result visualization module. The following shows the visualization results with the ‘SW’ substitution relationship, as follows:
(1) Single element visualization of gene sequences in five regions of China, Australia, Brazil, Colombia, and France, as shown in Fig. 3.



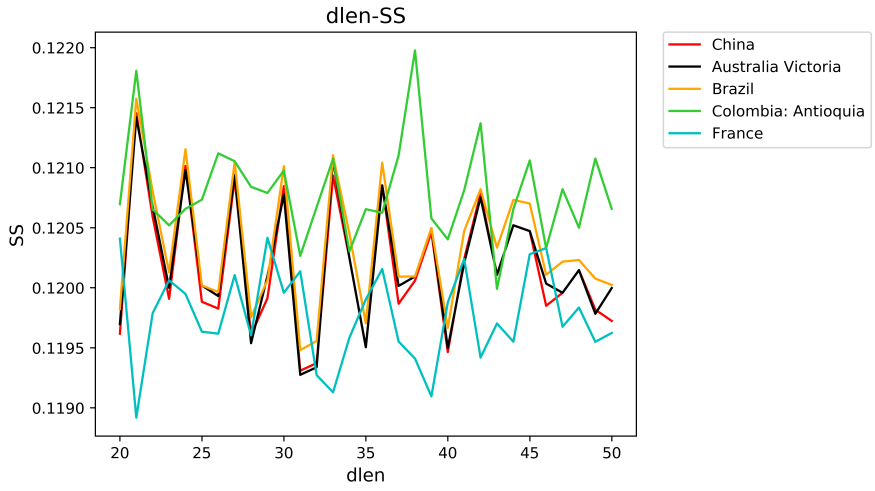
(a)



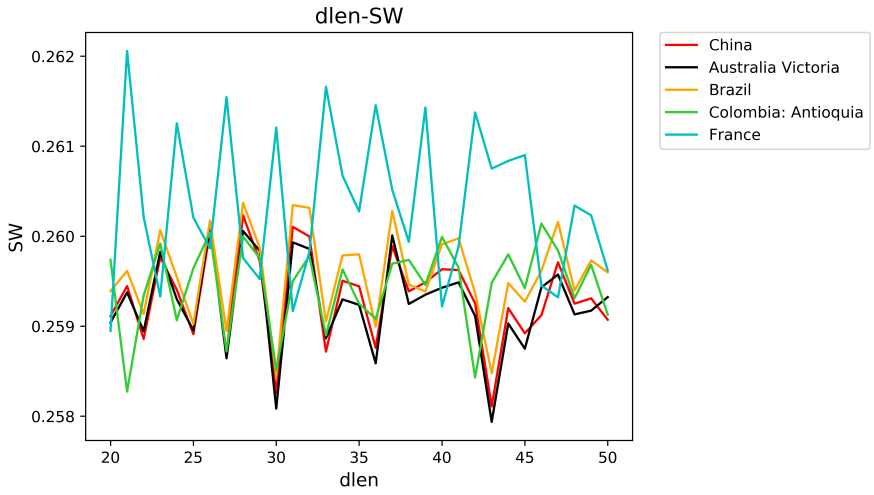
(b)

Fig. 3 Single-element Visualization of Gene Sequence

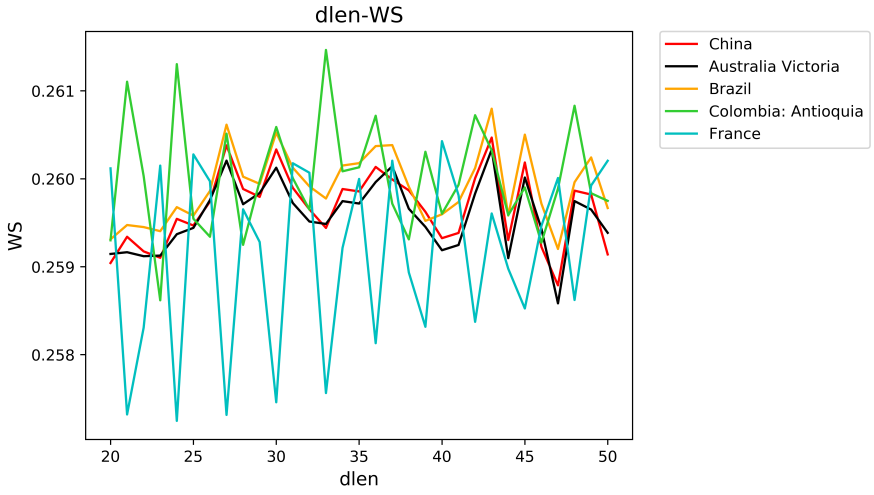
(2) Visualization of the location of gene sequence elements in five regions of China, Australia, Brazil, Colombia, and France, as shown in Fig. 4.



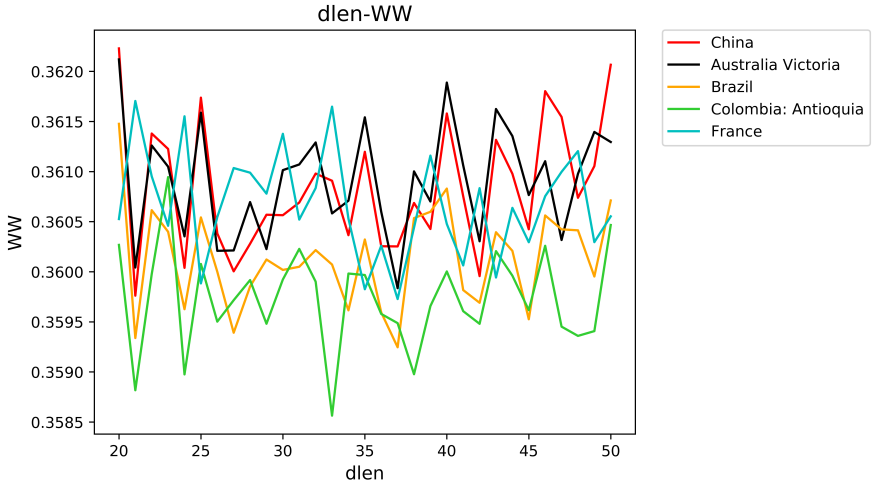
(a)



(b)



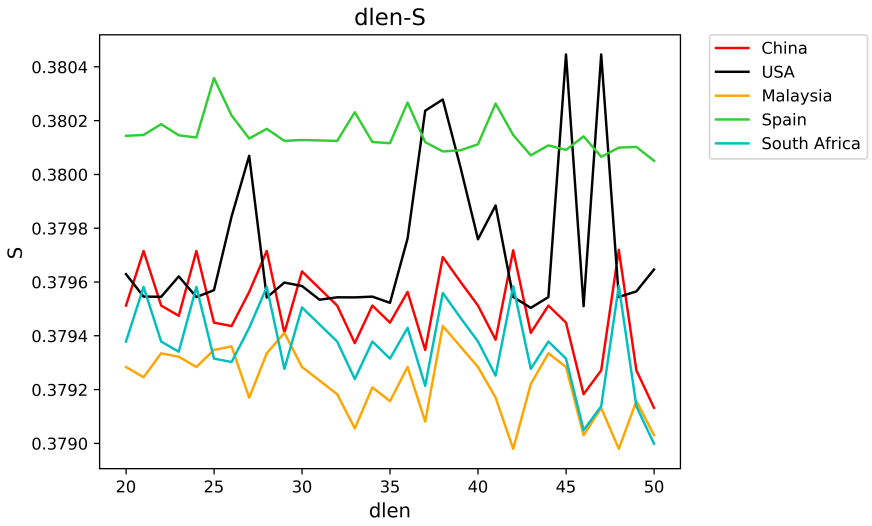
(c)



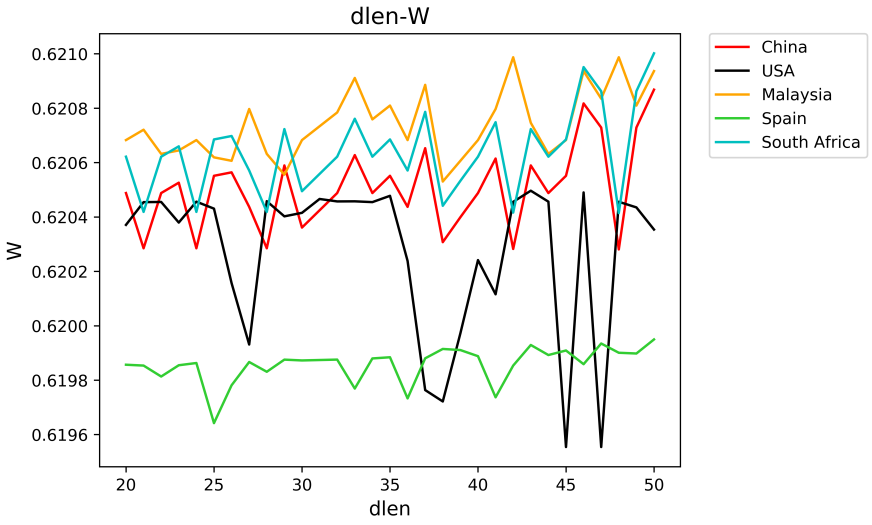
(d)

Fig. 4 Element-Position Visualization Of Gene Sequence

(3) Single element visualization of gene sequences in five regions of China, United States, Malaysia, Spain and South Africa, as shown in Fig. 5.



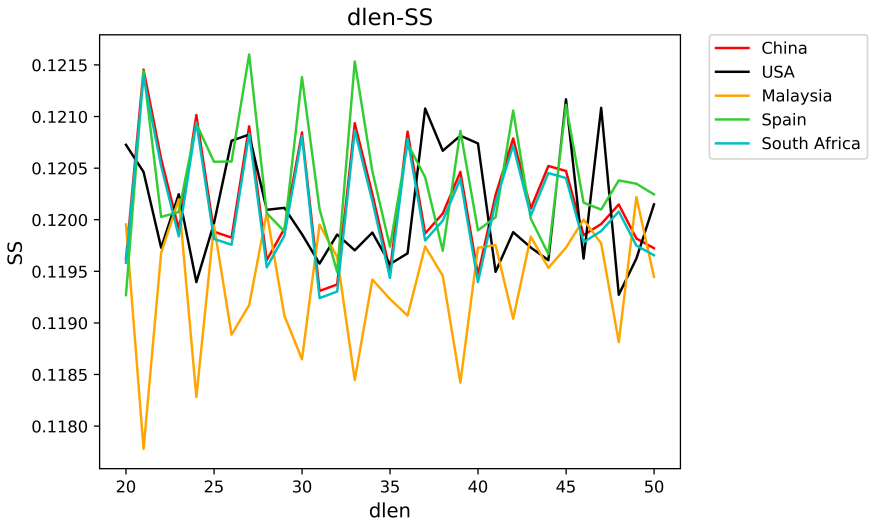
(a)



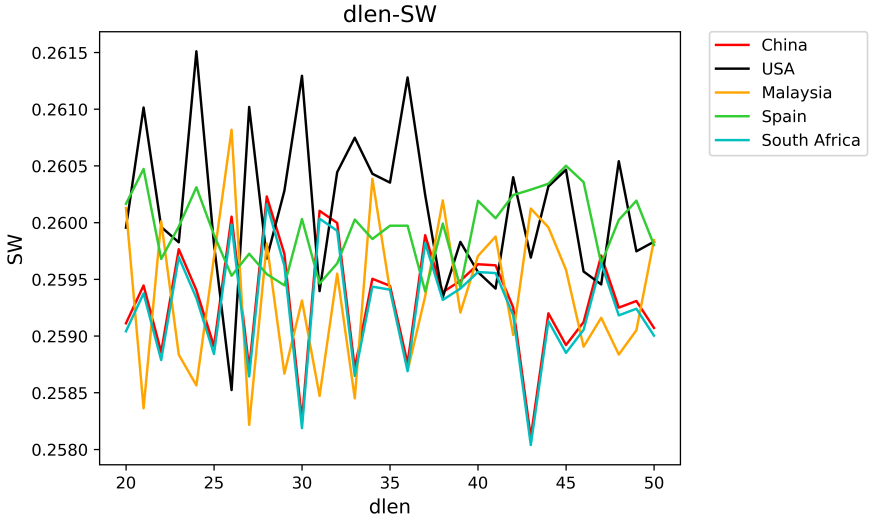
(b)

Fig. 5 Single-element Visualization of Gene Sequence

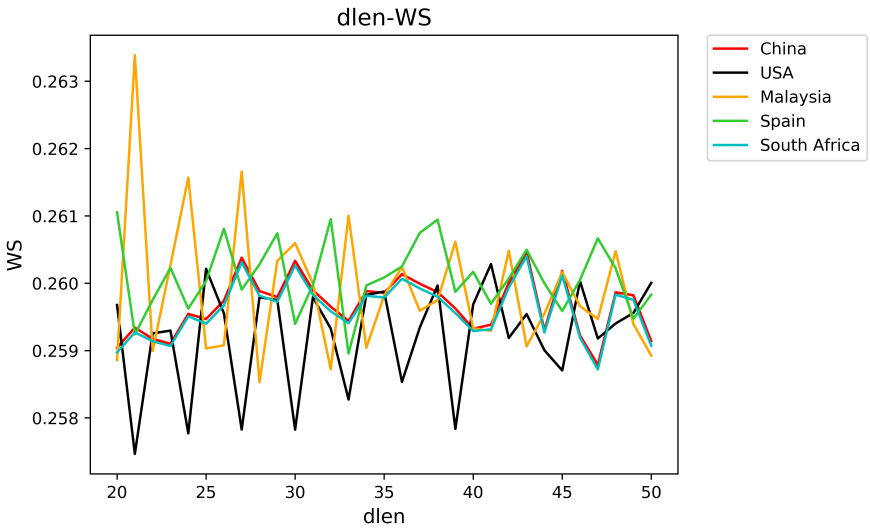
(4) Visualization of the location of gene sequence elements in five regions of China, United States, Malaysia, Spain and South Africa, as shown in Fig. 6.



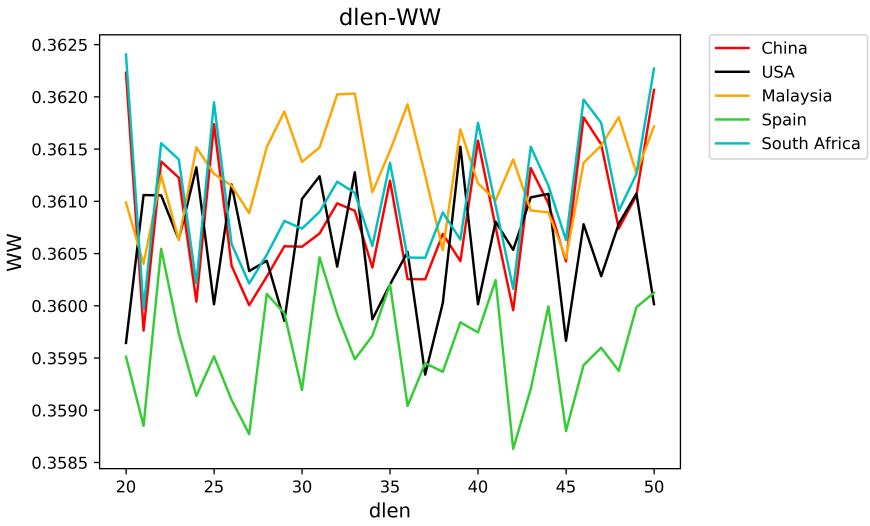
(a)



(b)



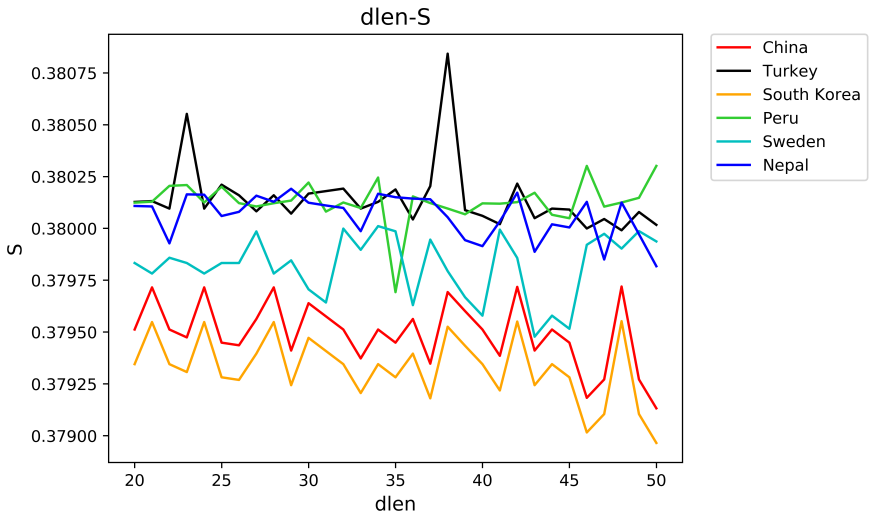
(c)



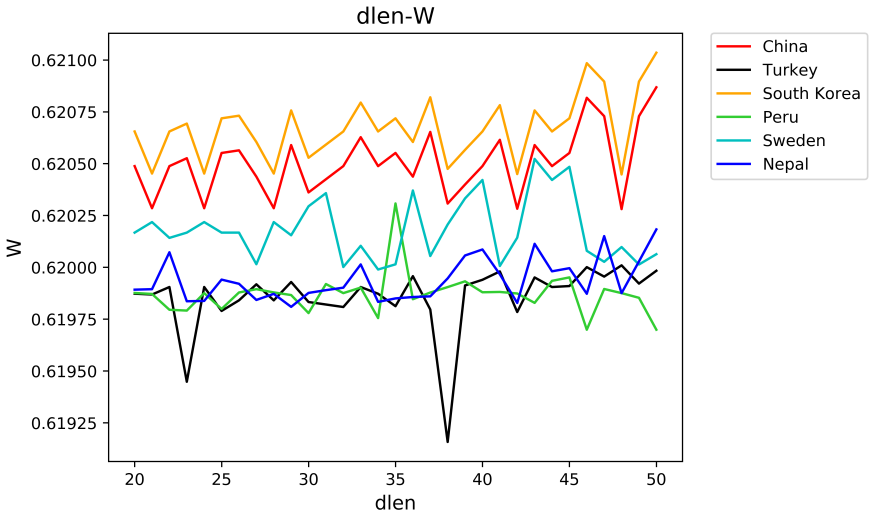
(d)

Fig. 6 Element-Position Visualization Of Gene Sequence

(5) Single element visualization of gene sequences in five regions of China, Turkey, South Korea, Peru, Sweden, and Nepal, as shown in Fig. 7.



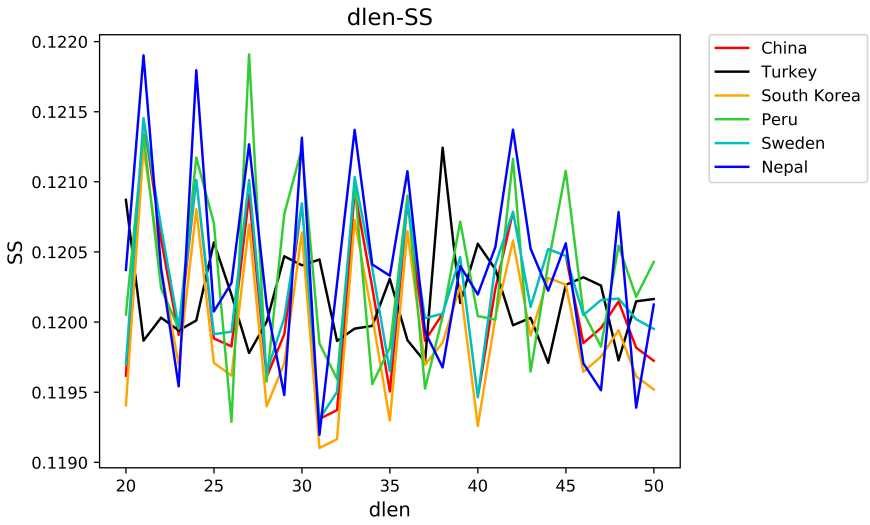
(a)



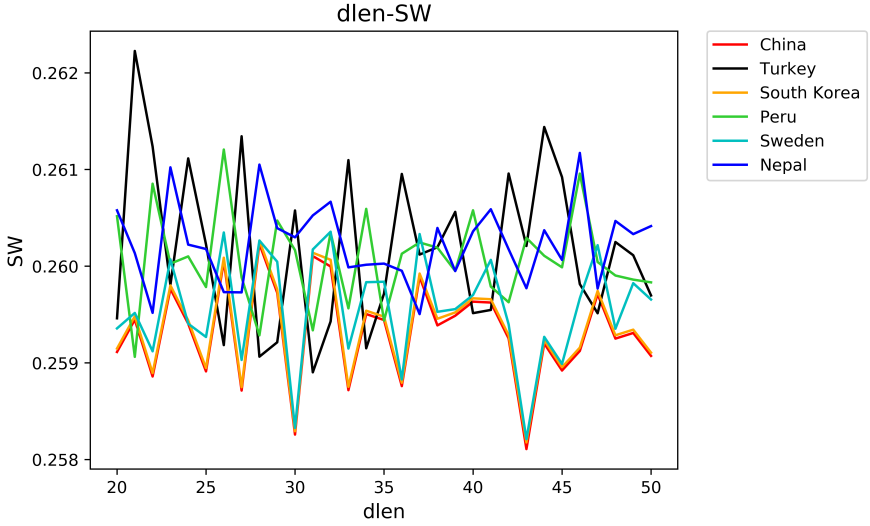
(b)

Fig. 7 Single-element Visualization of Gene Sequence

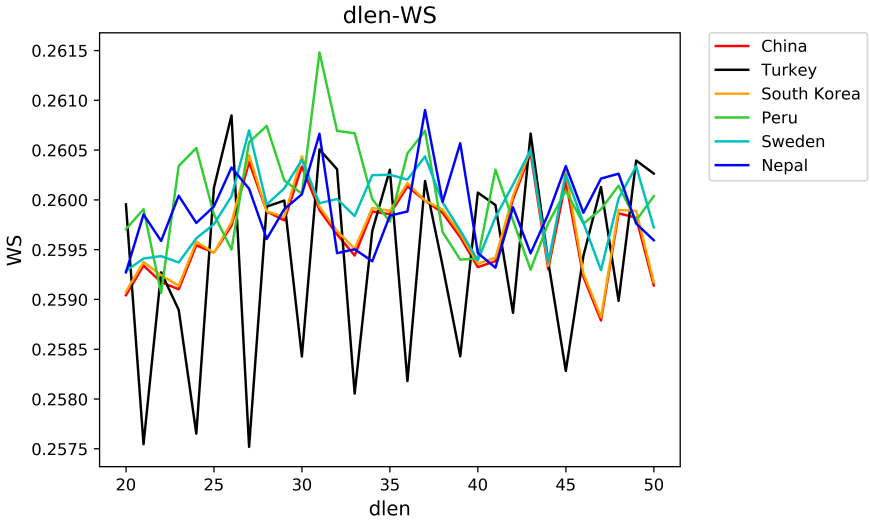
(6) Visualization of the location of gene sequence elements in five regions of China, Turkey, South Korea, Peru, Sweden, and Nepal, as shown in Fig. 8.


dlen-SS
SS
dlen
0.1220
0.1215
0.1210
0.1205
0.1200
0.1195
0.1190
20
25
30
35
40
45
50
China
Turkey
South Korea
Peru
Sweden
Nepal

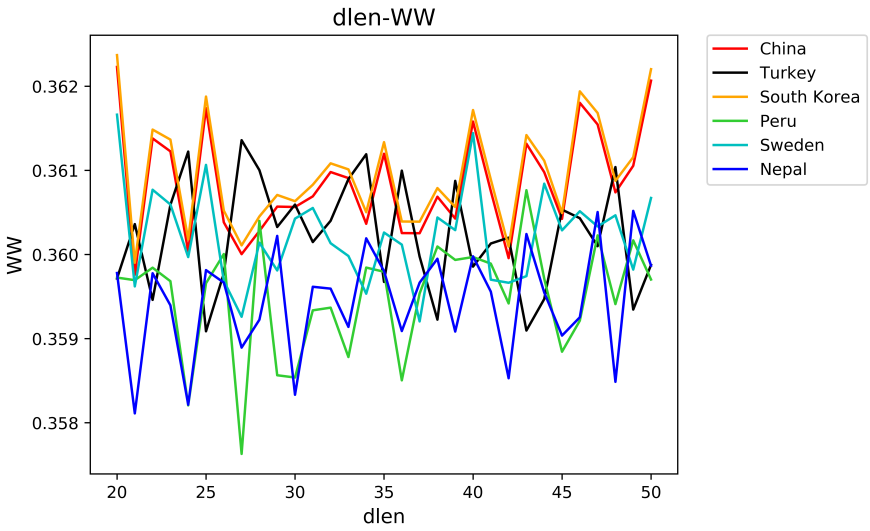
(a)


dlen-SW
SW
dlen
0.262
0.261
0.260
0.259
0.258
20
25
30
35
40
45
50
China
Turkey
South Korea
Peru
Sweden
Nepal

(b)


dlen-WS
WS
dlen
0.2615
0.2610
0.2605
0.2600
0.2595
0.2590
0.2585
0.2580
0.2575
20
25
30
35
40
45
50
China
Turkey
South Korea
Peru
Sweden
Nepal

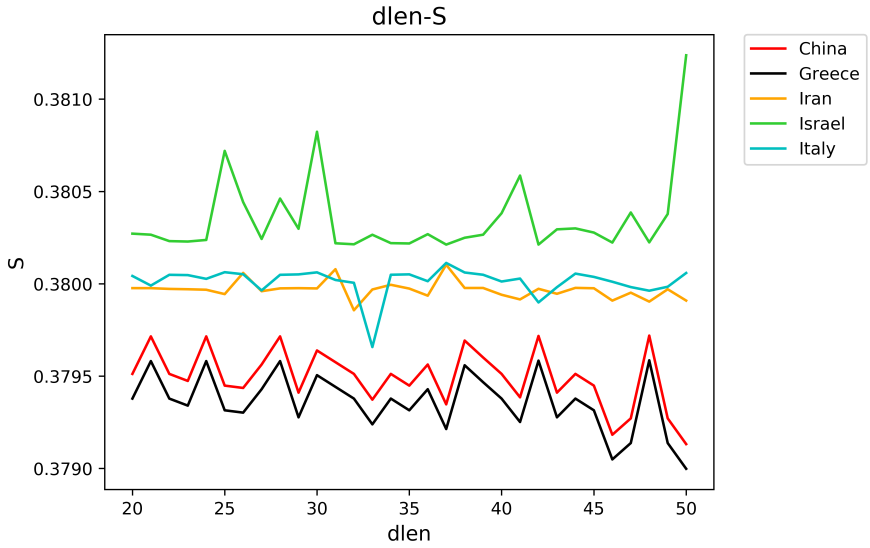
(c)



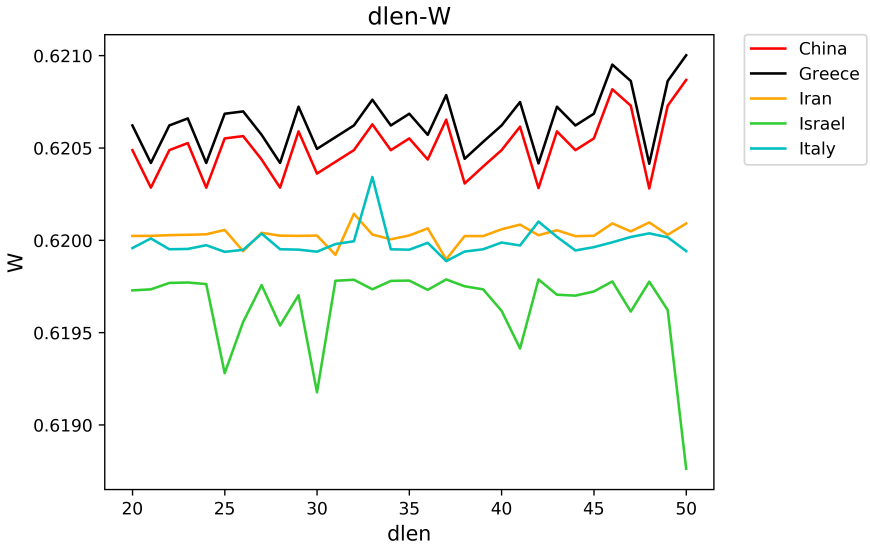
(d)

Fig. 8 Element-Position Visualization Of Gene Sequence

(7) Single element visualization of gene sequences in five regions of China, Greece, Iran, Israel, and Italy, as shown in Fig. 9.



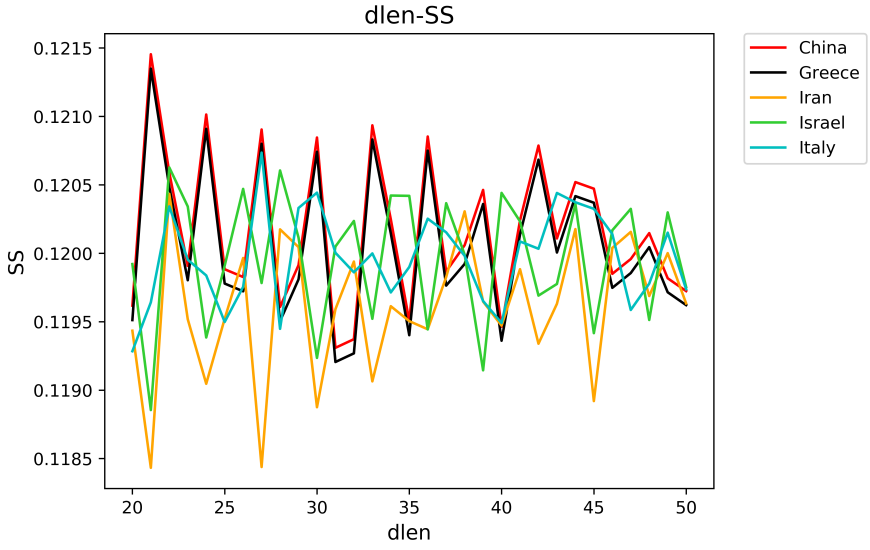
(a)



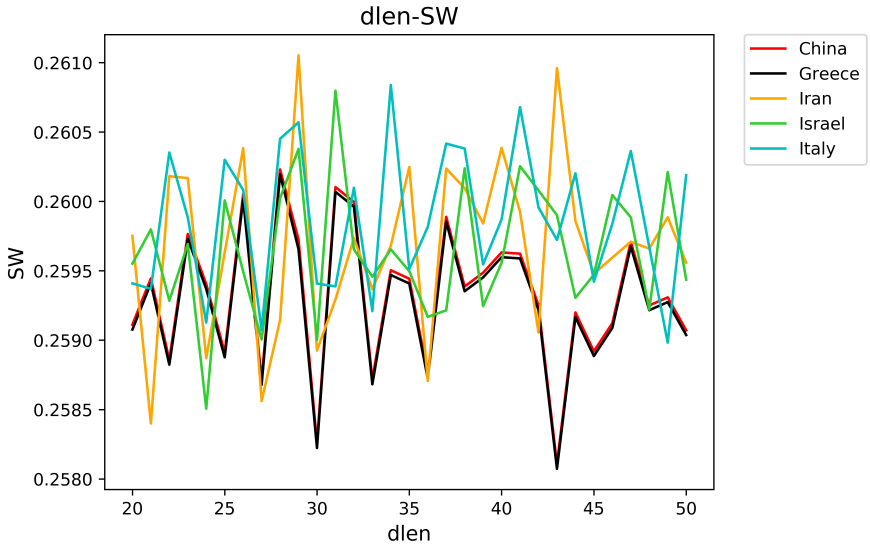
(b)

Fig. 9 Single-element Visualization of Gene Sequence

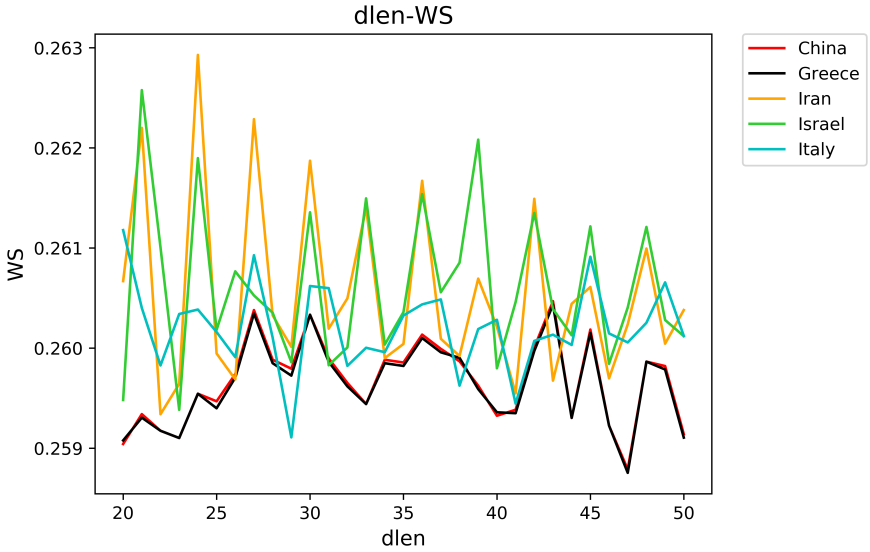
(8) Visualization of the location of gene sequence elements in five regions of China, Greece, Iran, Israel, and Italy, as shown in Fig. 10.



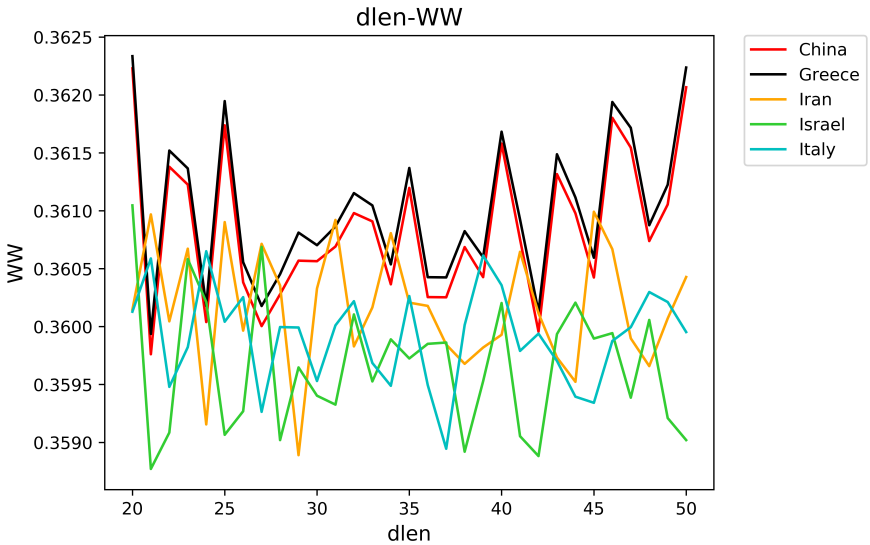
(a)



(b)



(c)



(d)

Fig. 10 Element-Position Visualization Of Gene Sequence

Discussion

The sequence of new coronavirus genes in 18 regions, in the four comparison groups, the sequence elements and sequence element position statistics measured by the ”SW” base substitution relationship have different distribution characteristics, and the virus genes in each region. The sequences show different similarities and differences. Different gene sequences have different comparison results in different sequence elements or sequence element position relationships. In the comprehensive analysis, compared with 18 regions, four regions with the most similar gene sequence to the sample data submitted by China are Australia. , South Africa, South Korea, Greece.

Although the gene sequences of each region is used as a unit in this article, only the ”SW” base substitution relationship is selected for display of results, but in fact, there may be other graphs showing the comparison results of base substitution relationships.

Conflict Interest

No conflict of interest has been claimed.

Acknowledgements The authors would like to thank NCBI,GISAID, CNGBdb for providing invaluable information on the newest dataset collections of SARS-CoV-2 to support this project working smoothly.

References

1. 陈嘉源，施劲松，丘栋安，刘畅，李鑫，赵强，阮吉寿，高山，*2019* 新型冠状病毒基因组的生物信息学分析, 生物信息学, 2020.
2. 郑智捷，郑昊航， 变值测量结构及其可视化统计分布, 光子学报, 2011.
3. ZHENG J, *Conditional Probability Statistical Distributions in Variant Measurement Simulations*, Acta Photonica Sinica, 2011.
4. ZHENG J, *Novel Pseudorandom Number Generation Using Variant Logic Framework[M]. Singapore: Variant Construction from Theoretical Foundation to Application*, Springer, 2019.
5. Zheng, J, *Variant Logic Construction under Permutation and Complementary Operations on Binary Logic. In: Zheng, J., Ed., Variant Construction from Theoretical Foundation to Applications*,Springer, 2019.
6. Zheng, J, *Variant Logic Construction under Permutation and Complementary Operations on Binary Logic. In: Zheng, J., Ed., Variant Construction from Theoretical Foundation to Applications*,Springer, 2019.
7. QY He et al, *Research on Functional Protein*, Science Press, 2012 (Chinese).
8. J. Barciszewski, V. Erdmann, *Noncoding RNAs*, Kluwer Academic Publishers, 2003 (Chinese).
9. FZ Song, Genomics, *Military Medical Science Press 2011 (Chinese)* 宋方洲，基因组学, 军事医学科学出版社, 2011(Chinese).