Comparative Study of Pathogenic Viruses Carried on Pairs of Species

Xin Zhang
Yunnan Uinversity
Zhaoyu Pan
Yunnan University
Jeffrey Zheng (conjugatelogic@yahoo.com)
Yunnan University

Research Article

Keywords: metagenomic analysis system MAS, variant theory, variant map, line chart, genetic mutation, variation

Posted Date: January 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-72028/v2

Comparative Study of Pathogenic Viruses Carried on Pairs of Species

Xin Zhang, Zhaoyu Pan, Jeffrey Zheng

Abstract The new coronavirus was checked on December 12, 2019, and spread rapidly over time. It has become a public health event spreading around the world. Until this time, the source of the virus remains controversial. In this paper, a series of SARS-CoV-2 genomes were collected using the A_1 module of the MAS for visualization. Pairs of genomes are compared under similarity relationships between SARS-CoV-2 and other deadly viruses carried by different species. Through the proposed method of variant construction, it provides important information to understand similarity properties among genomes. The comparison mechanism provides an efficient and fast similarity mode to compare with a whole genome at multiple levels of hierarchical measurements to provide variation information on internal correlation to a certain extent. Sample results are intuitively expressed through a list of 1D visual line charts for various distributions.

Keyword metagenomic analysis system MAS, variant theory, variant map, line chart, genetic mutation, variation

Jeffrey Zheng[1,2]
1,Key Laboratory of Quantum Information of Yunnan
2,Key Laboratory of Software Engineering of Yunnan
Yunnan University, Kunming, e-mail: conjugatelogic@yahoo.com

Xin Zhang
Yunnan University, e-mail: 752282264@qq.com

Zhaoyu Pan
Yunnan University, e-mail: 584844284@qq.com

Funding Supported by the NSFC (62041213), the Key Project of Quantum Communication Technology (2018ZI002)

Introductioon

At the beginning of 2020, the global pandemic of the new coronavirus cast a shadow over this new year, and countries around the world are actively responding and trying to overcome difficulties. However, there is still a huge controversy about the origin of the new coronavirus. It is an effective method through the combination of metagenomic [1-9] and sequence alignment.

Fifteen modules of {A,B,C} three groups in the metagenomic analysis system (MAS) provide unique capacities to support wider applications. This article shows the specific performance of the A_1 function module of the MAS in practical applications, and discusses the relationship between the deadly viruses carried by different species.

There are many comparison methods at this stage, such as the Needleman-Wunsh algorithm [10] and Smith-Waterman algorithm [11] based on alignment mode, and other alignment algorithms based on misalignment using k-mers [12] as the core . Through these algorithms, much research has been conducted on the source of the new coronavirus and the intermediate host, and some results have been achieved. However, these comparison methods have high complexity.

Through sequence alignment, we can judge the similarity between sequences and judge whether they are homologous sequences according to the degree of similarity [13]. Sequences with high similarity usually have a higher chance of being homologous sequences, and at the same time, homologous sequences usually have a higher similarity relationship. Although these are not necessarily true, they have correspondence in most cases, so judging homology by similarity is a way to be affirmed.

There are many challenges in finding homologous sequences and even viruses that may come from the middle array. It is very important to effectively screen and judge the virus sequences. Considering the special importance of Koch's Postulate in the period of genomics [14, 15], it is necessary to find proper techniques to resolve this type of difficulty. Facing a sea of biological data every where [16], it is really a top challenge to generate meaningful pictures emerged from those types of meaningless datasets.

In this paper, based on variant theory, combined with the random characteristics of gene sequences, the similarity relationship between deadly viruses carried by different species is analyzed. The differences in the direction of variation are compared to provide an efficient and fast similarity model, which can be compared with the entire genome at multiple levels of measurements, thereby providing internally related variation information to a certain extent new coronavirus research provides a new perspective.

Variant theory is based on classical logic [17], and variant mapping is performed by the variant logic function. Variant conversion, variant measurement, and variant projection form a complete set of variant measurement systems. In this paper, 1D visual line charts are used to carry out variant projection to show the correlation between sequences clearly and intuitively. Variant theory has achieved a series of achievements. In 2018, a monograph [18] was published on the basis of a phased

arrangement to introduce the system in detail and elaborate its application in various aspects.

Aim of The Study

By analyzing the similarity relationship between different sequences, the comparison result of the similarity relationship between the deadly virus carried by different species and the new coronavirus is obtained. By comparing and analyzing the different mutation results of the same virus, we can obtain the different characteristics of the virus when it is mutated, and at the same time provide the mutation information of internal correlation to a certain extent. The mutation pattern of the new coronavirus was explored.

Materials and Methods

The material uses the viral gene sequence downloaded from the NCBI and GISAID, and selects deadly viruses from bats, pangolins, and pigs that can infect humans and new coronavirus for comparison.

The core method used is a probability statistical model based on variant construction. Through segmentation and statistics, the comparison results are mapped to the 1D plane, and 1D line charts are drawn.

Input and Segment Statistics

The main function of this part is to separately count and save the number of four different bases in each segment of the two sequences involved in the comparison, and to provide a measurement basis for the next module.

Suppose the lengths of the two sequences S_1, S_2 participating in the alignment are L1 and L2 respectively, the length of the segment is m, N is the number of segments, then there are $S_1=a_1a_2...a_{N_1}$, $S_2=b_1b_2...b_{N_1}$, count the number of bases A, T, C, G in each segment in sequence $\{MA, MT, MC, MG\}$, and record.

Data Measurement

The main function of this part is to use the statistical values obtained in the segment statistics module to calculate the ratio of the bases in the corresponding positions of

the sequences S1 and S2, respectively, to establish a ratio set and to provide a data basis for the projection module.

Because the selected gene sequences are similar in length, but there are still personal differences, we choose $N = min(N1, N2)$ to format the longer sequence and delete the "excessive segments" at the end of the longer sequence to ensure participation in the comparison data. The length is the same (it affects the integrity of the sequence to a certain extent, but for the global alignment of the sequence, its impact can be ignored).

The entire calculation process is as follows:

$$\begin{cases} \mathrm{R_A}=\frac{\mathrm{M_A^{a1}}}{\mathrm{M_A^{b1}}},\frac{\mathrm{M_A^{a2}}}{\mathrm{M_A^{b2}}}\cdots\frac{\mathrm{M_A^{aN}}}{\mathrm{M_A^{bN}}} \\ \mathrm{R}_T=\frac{\mathrm{M}_T^{\mathrm{a1}}}{\mathrm{M}_T^{b1}},\frac{\mathrm{M}_T^{\mathrm{a2}}}{\mathrm{M}_T^{b2}}\cdots\frac{\mathrm{M}_T^{\mathrm{aN}}}{\mathrm{M}_T^{bN}} \\ \mathrm{R}_C=\frac{\mathrm{M}_C^{\mathrm{a1}}}{\mathrm{M}_C^{b1}},\frac{\mathrm{M}_C^{\mathrm{a2}}}{\mathrm{M}_C^{b2}}\cdots\frac{\mathrm{M}_C^{\mathrm{aN}}}{\mathrm{M}_C^{bN}} \\ \mathrm{R}_G=\frac{\mathrm{M}_G^{\mathrm{a1}}}{\mathrm{M}_G^{b1}},\frac{\mathrm{M}_G^{\mathrm{a2}}}{\mathrm{M}_G^{b2}}\cdots\frac{\mathrm{M}_G^{\mathrm{aN}}}{\mathrm{M}_G^{bN}} \end{cases}$$

Among them, $\mathrm{M_A^{a1}}$ represents the number of bases A in the $\mathrm{a_1}$ segment, and R_a represents the ratio of base A ratios formed by the ratio values of segments A and A in the two sequences corresponding to segments $\mathrm{S_1}$*and*$\mathrm{S_2}$

Visualization and Output

The main function of this part is to use the ratio data set obtained by the data measurement module to visualize it in the form of a line chart and analyze it through the image after output. Among them, the similarity analysis between sequences is mainly based on qualitative analysis of the tightness of entanglement between curves.

Figure 1 lists typical illustrations in different similar situations:
(a) This means that the two sequences are homologous and identical, and the characteristic is that the four curves representing different bases coincide into a straight line parallel to the X axis and the ordinate is 1;
(b) This means that the nonhomologous similarity between the two sequences has a large difference, and the characteristic is that the sparse and disordered overall fluctuation is large;
(c) This means that the two sequences are homologous and extremely similar, and they are characterized by tight entanglement and little overall fluctuation;
(d) This means that the two sequences have the possibility of homology but the similarity is relatively low. The characteristics of the two sequences are that they are tightly entangled, but at the same time have certain fluctuation characteristics..

Results and Discussions

Result

Figure 2 and Figure 3 show the comparison results of viruses collected from different species and the comparison results of viruses collected from different individuals of the same species.

Dicussions

Figure 2 shows that there are obviously different similarities between different viruses.

Analyzing Figure 2 (a)-(c), it can be observed that SARS-CoV-2 has a very high similarity with the virus carried by pangolin. This is because the sample here is a variant coronavirus collected from pangolin, but it only has a high similarity with the SARS carried in bat, but it shows a significant difference compared with PDcov.

Analysis of Figure 2 (d)-(e) shows that the SARS and PDcov carried by pangolin and bat have similar performance as SARS-CoV-2.

Analyzing Figure 2 (c), (e), and (f), it can be observed that PDcov has a low similarity with viruses carried by the other three species.

Analysis of Figure 3 (a) (b) (c) shows that (a) SARS-CoV-2 in different countries has a variation at the overall level, but the variation distribution is more uniform; (b) although the overall level of SARS virus collected in bat, there are variations on the above, the high variation is mainly concentrated in a small part of the front and tail of the sequence; (c) The variation of PDcov from pigs is mainly concentrated and part of the site, which is shown as a unique small protrusion.

Conclusion

Through experimental comparison, it is found that SARS-CoV-2 has a high homology relationship with SARS virus, but does not have a significant homology relationship with PDcov.

There are obvious differences in the mutation modes of the three viruses. From this point of view, the distribution of mutation sites can be observed on the basis of sequences with high homology to further determine the relationship between them, and can be used to find more reliable intermediate hosts.

Conflict Interest

No conflict of interest has been claimed.

Acknowledgements: Thanks to NCBI,GISAID,CNGBdb, Nextstrain and Zhigang Zhang to provide invaluable information on the newest dataset collections of SARS-CoV-2 and other coronavirus genomes,which promoted the smooth progress of this project.

References

1. C, Saccone and G. Pesole. Handbook of Comparative Genomics, John Wiley and Sons Inc. 2003 C.萨科内，G.佩索莱， 比较基因组学手册，化学工业出版社 2008
2. S Klusmann, The Aptamer Handbook: Functional Oligonucleotides and Their Applications, John Wiley and Sons Inc. 2005 斯文.克卢斯曼，核酸适配体手册，化学工业出版社 2013
3. 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
4. ZH Yang, Computational Molecular Evolution, Fudan University Press 2008 (Chinese)杨子恒，计算分子进化，复旦大学出版社 2008
5. BL Hao, Chaos and Fractals, Shanghai Science and Technology Press 2015 (Chinese) 郝柏林，混沌与分形， 上海科学技术出版社 2015
6. Alexander Isaev, Introduction to Mathematical Methods in Bioinformatics, Springer-Verlag 2006 生物信息学中的数学方法引论， 科学出版社 2011
7. Michael Yarus, Life From An RNA World, Harvard University Press 2010
8. J. Collado-Vides & R. Hofestadt, Gene Regulation and Metabolism, The MIT Press 2001
9. Y Huang, Generating Science of Molecular System, Science Press 2012 (Chinese) 黄原，分子系统发生学，科学出版社 2012
10. Needleman Saul B.,Wunsch Christian D.. A general method applicable to the search for similarities in the amino acid sequence of two proteins[J]. Academic Press,1970,48(3).
11. Smith T F,Waterman M S. Identification of common molecular subsequences.[J]. Journal of molecular biology,1981,147(1).
12. B. Edwin Blaisdell. A Measure of the Similarity of Sets of Sequences not Requiring Sequence Alignment. 1986, 83(14):5155-5159.
13. JX Yan, Koch ' s Postulate in the Period of Genomics, Universal Science 2013 (Chinese) 严家新基因组时代的科赫法则，环球科学 2013 https://huanqiukexue.com/plus/view.php?aid=23170
14. Dan E.Krane,Mi chae I L Raymer.生物信息学概论[M].清华大学出版社，2004
15. D.N.FREDRICKS and D.A.RELMAN,Sequence-Based Indentification of Microbial Patho · genes: a Reconsideration of Koch ' s Postulates, CLINICAL MICROBIOLOGY REVIEWS, 9(1) 1996 18-33 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC172879/pdf/090018.pdf
16. E Pennisi, Problem 17: How Will Big Pictures Emerge from a Sea of Biological Data? , Science Vol 309:94 in SCIENCE Magazine: Top 125 Scientific Problems http://science.sciencemag.org/content/sci/309/5731/78.2.full.pdf Science公布全世界最前沿125个科学问题：(问题 17: 怎样从海量生物数据中产生大的可视化图片？)
17. 欧文 · M · 柯匹 ,卡尔 · 科恩,Irving M.Copi,等. 逻辑学导论 [M]. 中国人民大学出版社
18. Jeffrey Zheng, "Variant Construction from Theoretical Foundation to Applications". Springer Nature 2019.https://www.springer.com/in/book/9789811322815.

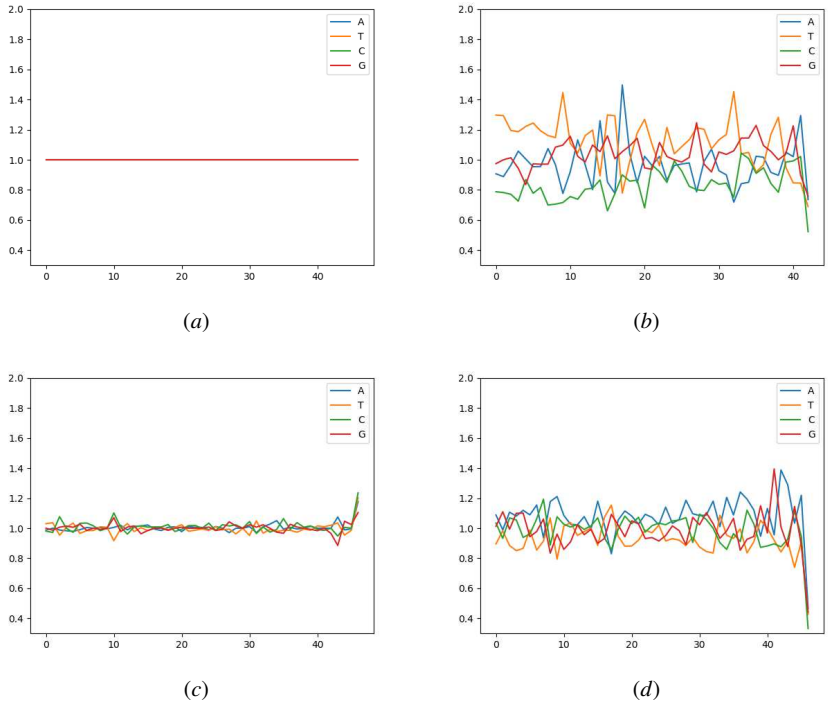


Fig. 1 Typical illustrations in different similar situations

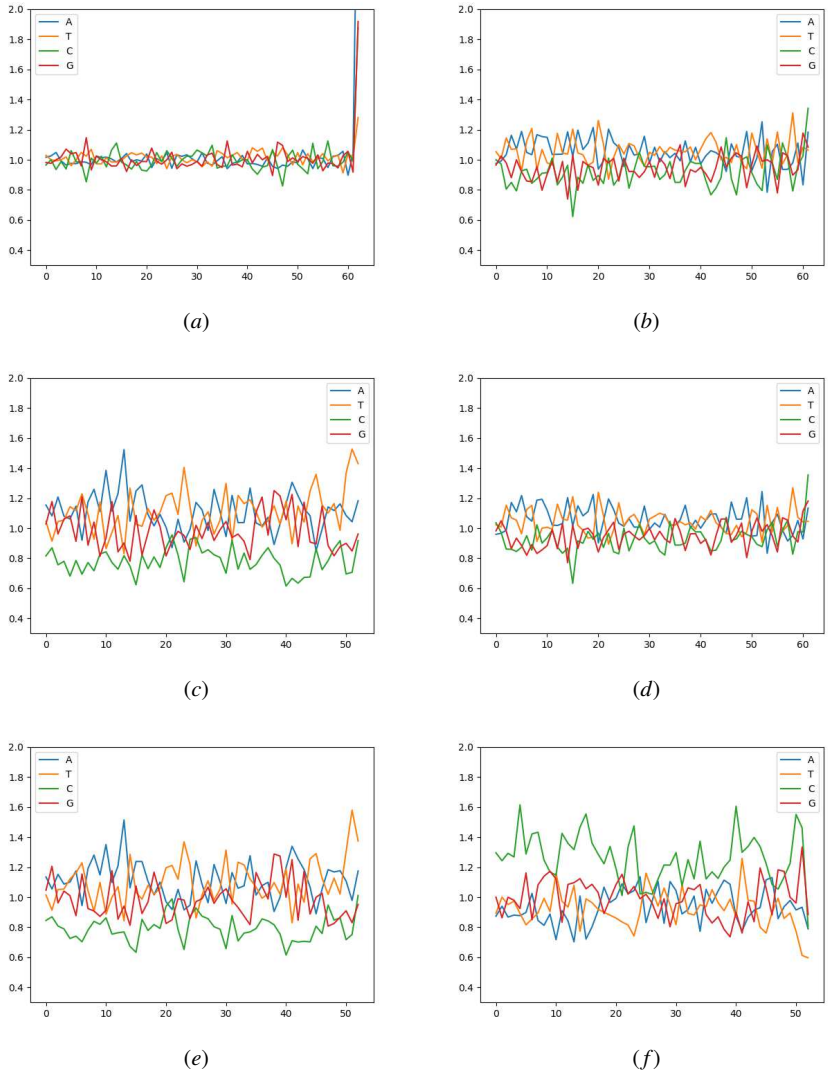
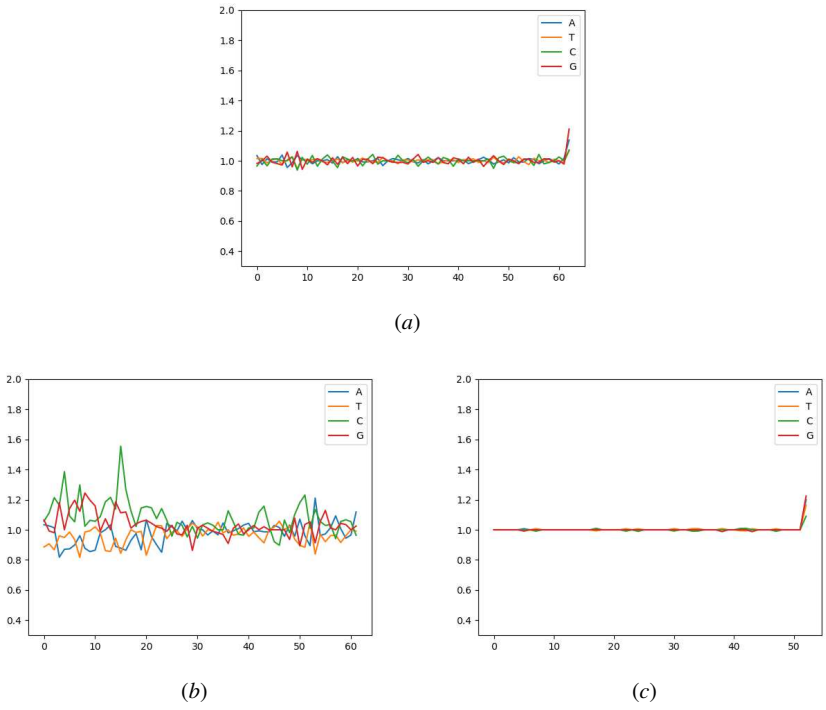


Fig. 2 (*a*)SARS-CoV-2 and pangolin virus;(*b*)SARS-CoV-2 and Bat-SARS-like RsSHC014 ;(*c*)SARS-CoV-2 and PDCov;(*d*)pangolin virus and Bat-SARS-like RsSHC014;(*e*)pangolin virus and PDCov;(*f*)Bat-SARS-like RsSHC014 and PDCov



(*a*)

(*b*) (*c*)

Fig. 3 (*a*)SARS-CoV-2-Wuhan and SARS-CoV-2-Canada;(*b*)Bat SARS-like RsSHC014 and bat-SL-CovZC45;(*c*)PDCov-KX022602 and PDCov-KX022605.

Figures

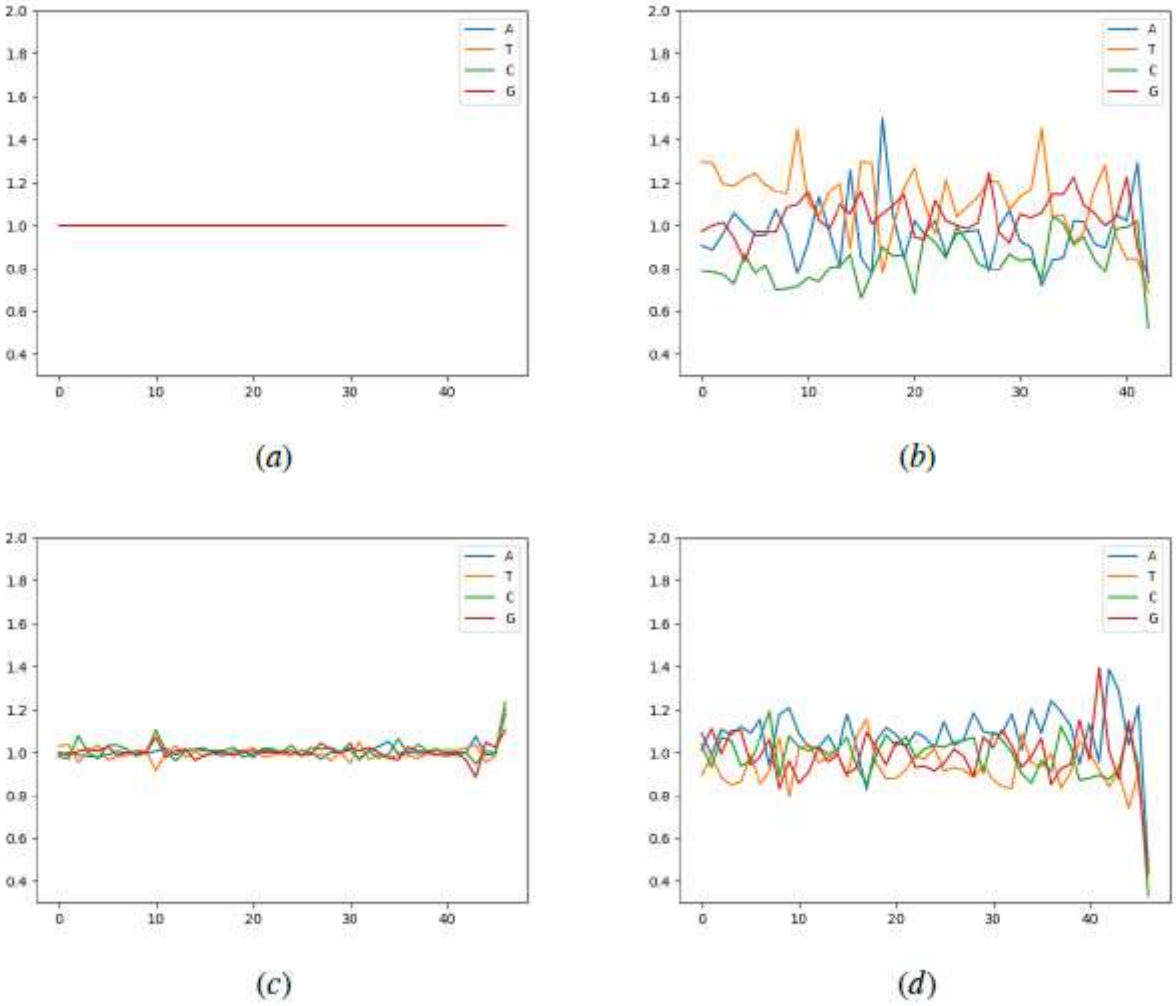


Figure 1

Typical illustrations in different similar situations

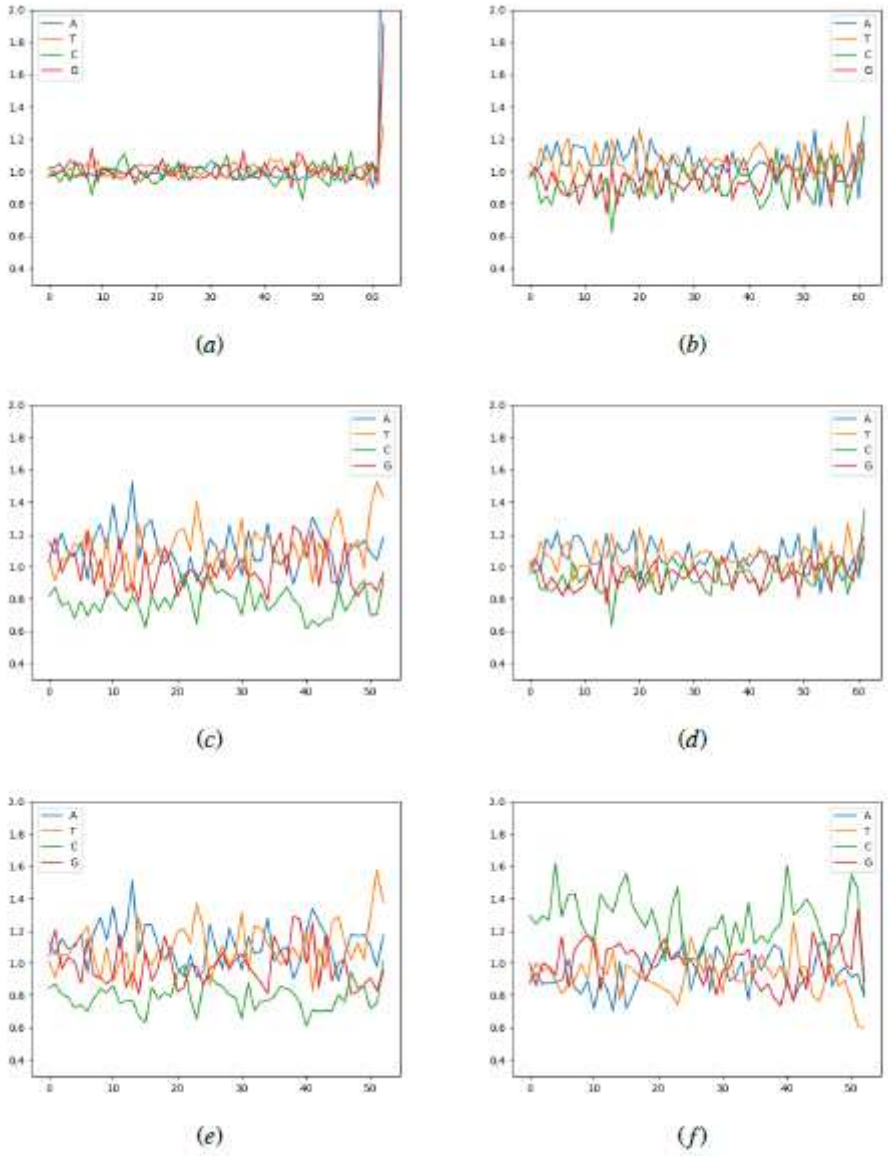


Figure 2

(a)SARS-CoV-2 and pangolin virus;(b)SARS-CoV-2 and Bat-SARS-like RsSHC014 ;(c)SARS-CoV-2 and PDCov;(d)pangolin virus and Bat-SARS-like RsSHC014;(e)pangolin virus and PDCov;(f)Bat-SARS-like RsSHC014 and PDCov

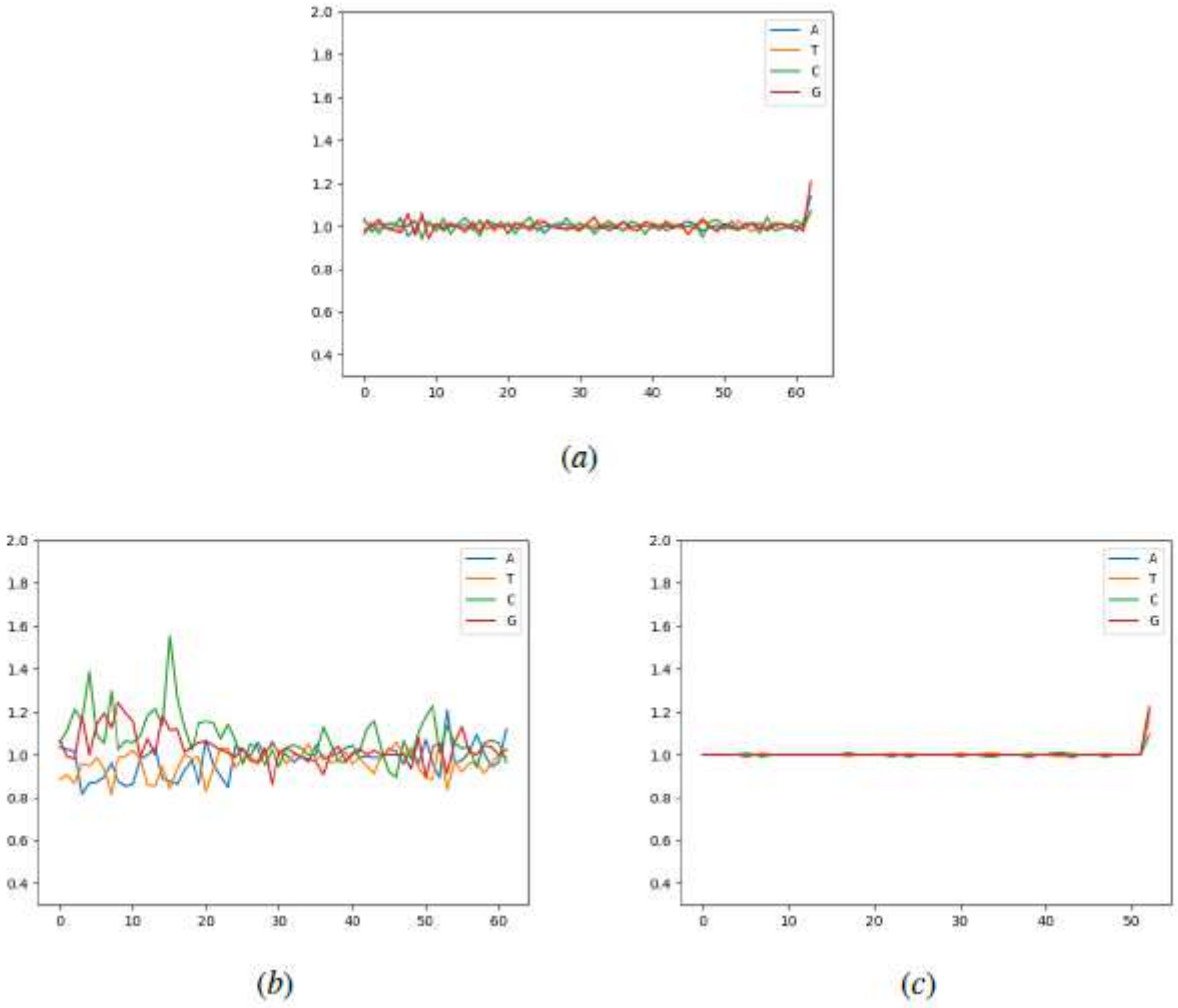


Figure 3

a)SARS-CoV-2-Wuhan and SARS-CoV-2-Canada;(b)Bat SARS-like RsSHC014 and bat- SL-CovZC45; (c)PDCov-KX022602 and PDCov-KX022605.