Viral Evolution of Multiple Coronavirus Genomes on Genomic Index Maps

Jeffrey Zheng[1,2,3,*], Minghan Zhu[3]

[1]Key Laboratory of Quantum Information of Yunnan, Yunnan University, Kunming 650900, China
[2]Key Laboratory of Software Engineering of Yunnan, Yunnan University, Kunming 650900, China
[3]School of Software, Yunnan University, Kunming 650900, China

*Corresponding author: E-mail: conjugatelogic@yahoo.com

DOI: 10.5185/amlett.2021.021600

Motivation: Multiple coronavirus genomes are normally organized by genomic information in a phylogenic tree to illustrate evolutionary variations. A novel scheme is represented to arrange various coronavirus genomes by two genomic indexes as 2D maps.

Results: For a genome, two unique invariants of genomic indexes provide an absolute position on a 2D region in real measurements. Clustering effects are provided complementary from Phylogeny technology. Samples of seventeen coronavirus and twenty-six SARS-CoV-2 genomes from various countries are selected. This provides an efficient scheme to identify variations of SARS-CoV-2 strains for better identifications on both complicated coronavirus clusters and SARS-CoV-2 group of viral evolution in intermediate and original hosts on selective environments.

Introduction

The outbreak of SARS-CoV-2 caused COVID-19 to start in December 2019 and is now pendamic. To the date of 22 June 2020, there are more than 8.84 million confirmed cases and 0.46 million deaths worldwide. Since SARS-CoV-2 is a specific coronavirus, it is extremely important to investigate the relationship among multiple coronaviruses to better understand COVID-19. The identification of the first coronavirus "infectious bronchitis virus (IBV) isolated from birds. Many coronaviruses have been identified from various animals: bat, camel, cat, dog, pig, murine, pangolin and whale. They may cause wider diseases with different levels of severity in a variety of hosts. It is essential to investigate proximal origin [**1**], probable intermediate hosts [**2**], evolutionary conservation [**3,4**], genetic diversity of Bat [**5-7**], genome-based classification/phylogenies [**8,9**], genetic evolution analysis [**10**] and coronavirus mutations affecting deadliness of strains [**11**] for medical treatments of COVID-19 patients.

Various coronavirus

Coronaviruses have positive-sense single-stranded RNAs, and their genomic size is 26 to 32 kilobases, the largest for an RNA virus [**12**]. The viruses appear crown-shaped under electron microscopy. Coronaviruses can be further divided into four groups [**1**]: alpha, beta, gamma, and delta corona virus, based on phylogenetic relationships and genomic structures. Coronaviruses occasionally jump across host barriers, often with lethal consequences. Alpha and beta coronaviruses only infect mammals and usually cause respiratory illness in humans and gastroenteritis in animals. Gamma and delta coronaviruses mainly infect birds, and no human infection has been reported. Six coronavirus-infected humans are 229E, NL63 (genus Alpha), OC43, HKU1, SARS-CoV, and MERS-CoV (Beta). SARS and MERS-CoV have caused large worldwide outbreaks with fatality, while others usually cause mild upper respiratory tract illnesses. SARS-CoV-2 was previously identified in a pneumonia patient [**10**] on 9 January 2020 to represent the seventh human-infecting coronavirus.

Genomic CpG deficiency

From a visual viewpoint [**13-15**] associated with CG proportions [**16**], genomic CpG deficiency provides an interesting scheme to illustrate multiple coronaviruses in 2D maps. From the represented maps, this research identified the SARS-CoV-2 genomes with the lowest measurement of CpG deficiency. From a measuring viewpoint, both CG proportions and CpG deficiency are probability/frequency measurements composed of rational numbers. This type of measurement based on the lengths of genomes makes 2D maps with maximal visual limitations. In addition, CG proportions cannot provide refined measurements on combinatorial variations of genomes in general.

Approach - genomic index map

Different entropy quantities were discussed: thermodynamics [**17**], hash-based re-ordering [**18**], entropy coding [**19**], hierarchical framework [**20**], combinatorial / mean / integrated / topological genomic indexes [**21,22**]. The genomic index provides unique identification for each genome to be invariant under given conditions. Based on these types of global quantitative characteristics, it is convenient for large numbers of genomes to be located in a certain geometric region to be collected as clusters.

Three workflows are involved.

1. Vector of Genome → Four Probability Vectors → Sixteen Probability Vectors → Entropy → Sixteen Indexes → Mean → A Mean Index (MI)
2. Vector of Genome→Four Probability Vectors→ Sixteen Probability Vectors→Mean→ One Probability Vector→ Entropy→ An Integrated Index (II)
3. {(MI, II)}→ Mapping→ A Genomic Index Map

Methods - Entropy measurement

Let a vector Z with (m+1) elements,

$$Z = (Z_0, Z_1, \dots, Z_j, \dots, Z_m), 0 \le Z_j \le M \text{ and}$$
$$M = \sum_{j=0}^{m} Z_j.$$

Under this condition, let $P_j = \frac{Z_j}{M}$ be the j-th probability measurement, and a relevant information entropy Z can be determined and restricted in a $[0, log_2(m+1)]$ region.

$$eZ = -\sum_{j=0}^{m} P_j log_2(P_j), eZ \in [0, log_2(m+1)]$$

$$1 = \sum_{j=0}^{m} P_j$$

2D (MI, II) Indexes A pair of indexes corresponds to $eZ(MI, II) = (eZ_{MI}, eZ_{II})$. There is one 2D position determined by the genome Z in the square on the $[0, log_2(m+1)] \times [0, log_2(m+1)]$ region.

Multiple genomes

For multiple genomes $\{Z^t\}, 1 \le t \le T$ on maximal T members of each (MI, II) projection, a total number of T positions can be collected on a 2D square of $\forall(eZ_{MI}^t, eZ_{II}^t), 1 \le t \le T$. This provides a special distribution for whole genomes of Tmembers on (MI, II) projection based on a pair of entropy measurements.

$$EZ_{MI,II} = \sum_{t=1}^{T} eZ_{MI,II}^t = \sum_{t=1}^{T} (eZ_{MI}^t, eZ_{II}^t)$$

Each $EZ_{MI,II}$ represents an index map corresponding to a $[0, log_2(m+1)] \times [0, log_2(m+1)]$ region.

Datasets

A total of 43 genomes are selected and shown in **Table 1** with twenty-six genomes from 26 countries/regions for COVID-19 and seventeen genomes for other coronavirus genomes. Special considerations are carried out in balancing multiple coronavirus species and typical COVID-19 regions with distinct properties.

Six coronavirus genomes: Scotophilus bat_NC_009657, Eidolon bat_HQ_728482, Infectious bronchitis virus_MN517817, Duck coronavirus_ KM454473, Mink coronavirus_MT457401, Murine murine_AC_000192; three human coronaviruses: HCOV-229E NC-002645, HCOV-HKU1 NC-006577, HCOV-OC43 NC-006213; three deadly coronaviruses: SARS AY-508724, MERS JX-869059, EBOLA NC-002549; three animal coronaviruses: PDCOV KX-022602, Bat SL-CovZC45, Pangolin MT-084071 and two other viruses: H1N1 and H3N2. SARS-CoV-2 genomes are selected from twenty-six countries from Australia to Wales.

Table 1. Forty-three Coronavirus genomes and Twenty-six genomes of COVID-19 collected from 26 countries/regions.

No.	Sequence No.	Mean Index	Integrated Index	Others
hCoV-19/Australia	EPI_ISL_407893	2.536	4.0004	2020-01-24
hCoV-19/Belgium	EPI_ISL_407976	2.5384	4.0007	2020-02-03
hCoV-19/Brazil	EPI_ISL_412964	2.5394	4.0009	2020-02-25
hCoV-19/Canada	EPI_ISL_413014	2.5381	4.0008	2020-01-25
hCoV-19/Wuhan	EPI_ISL_402119	2.5394	4.0008	2019-12-30
hCoV-19/Denmark	EPI_ISL_415646	2.5369	4.0007	2020-03-03
hCoV-19/England	EPI_ISL_407071	2.5369	4.0004	2020-01-29
hCoV-19/France	EPI_ISL_406596	2.5368	4.0004	2020-01-23
hCoV-19/Germany	EPI_ISL_406862	2.5363	4.0004	2020-01-28
hCoV-19/ HongKong	EPI_ISL_412028	2.5367	4.0004	2020-01-22
hCoV-19/Iceland	EPI_ISL_417481	2.5357	4.0005	2020-03-13
hCoV-19/India	EPI_ISL_413522	2.5365	4.0007	2020-01-27
hCoV-19/Italy	EPI_ISL_410545	2.5363	4.0005	2020-01-29
hCoV-19/Japan	EPI_ISL_408665	2.5368	4.0006	2020-01-29
hCoV-19/ Luxembourg	EPI_ISL_413593	2.5358	4.0005	2020-02-29
hCoV-19/Netherlands	EPI_ISL_413564	2.5359	4.0003	2020-03-01
hCoV-19/Portugal	EPI_ISL_413647	2.5362	4.0006	2020-03-01
hCoV-19/Russia	EPI_ISL_415710	2.5357	4.0006	2020-03-15
hCoV-19/Scotland	EPI_ISL_414027	2.5363	4.0004	2020-03-04
hCoV-19/Singapore	EPI_ISL_406973	2.5366	4.0005	2020-01-23
hCoV-19/Spain	EPI_ISL_414598	2.5356	4.0004	2020-03-05
hCoV-19/Sweden	EPI_ISL_411951	2.5365	4.0005	2020-02-07
hCoV-19/Switzerland	EPI_ISL_413019	2.5374	4.0007	2020-02-26
hCoV-19/Taiwan	EPI_ISL_406031	2.5368	4.0006	2020-01-23
hCoV-19/USA	EPI_ISL_404253	2.538	4.0008	2020-01-21
hCoV-19/Wales	EPI_ISL_413555	2.5361	4.0004	2020-02-27
Duck coronavirus	KM454473	2.5697	4.0038	
Eidolon bat	HQ_728482	2.5156	3.9914	
Infectious bronchitis virus	MN517817	2.5483	4.0006	
mink coronavirus	MT457401	2.5577	3.9883	
Murine	AC_000192	2.5266	3.9969	
Scotophilus bat	NC_009657	2.5397	4.0009	
bat-SL-CovZC45		2.5452	3.9982	
Ebola	NC_002549	2.6077	3.9999	
H1N1_RUSSIA	sequence_395068	2.5469	3.9996	
H3N2		2.554	3.9991	
HCoV-229E	NC_27317	2.5205	4.0054	
HCoV-HUK1	NC_29913	2.4983	4.0315	
HCoV-NL63	NC_27553	2.4922	4.0251	
MERS-CoV	NC_30090	2.5092	3.9868	
PDCoV	KX_022602	2.515	3.9757	
SARS	AY_485277	2.5207	3.9866	
Pangolin	MT_084071	2.5295	3.9988	

Sample results

All 43 genomes are shown in **Fig. 1(a)** as a regular map, and an enlarged map is shown in **Fig. 2**. Twenty-six SARS-CoV-2 genomes plus a Scotophilus bat genome and the center (black +) are shown in **Fig. 1(b)** as an enlarged map 100-fold from **Fig. 1(a)** & **Fig. 2**.

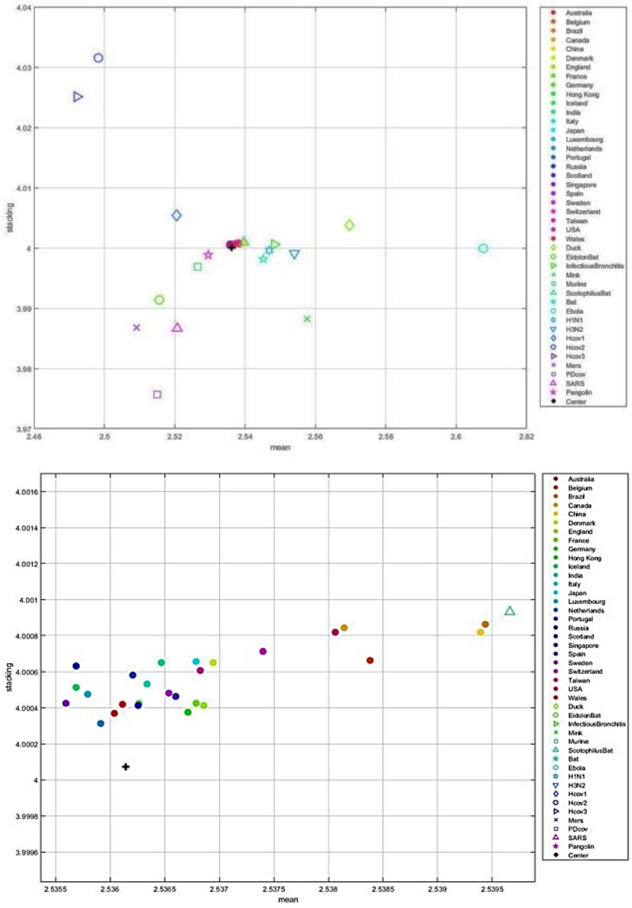

Fig. 1. Coronaviruses on Genomic Index Maps m= 16 (a) Forty-Three Coronaviruses (b) Twenty-Seven Genomes of SARS-CoV-2, Scotophilus Bat plus Center.

Discussion

In **Fig. 1(a)** and **Fig. 2**, twenty-six genomes of SARS-CoV-2 are formed as a cluster at (2.5365, 4.0005) near the center position, the left side of the center is a red star to show a Pangolin genome near the SARS-CoV-2 region, and the left-down position is a green square to show a murine genome. Further left-down positions have four genomes: a green cycle (Eidolon Bat), a deep-blue cross (MERS), a red triangle (SARS) and a red square (PDcov). The left-up position is a blue rhombus (Hcov1). Further left-up positions have a blue triangle (Hcov3) and a blue cycle (Hcov2). The right side of the center is a Scotophilus bat genome. On further right locations, three genomes are Bat (a blue star), H1N1 (a six-fold star) and infectious bronchitis virus (a blue triangle). Three genomes are located on the left side: a blue delta (H3N2), a yellow rhombus (duck) and a blue cycle (Ebola). A mink genome (a blue cross) is located on the right-down side.

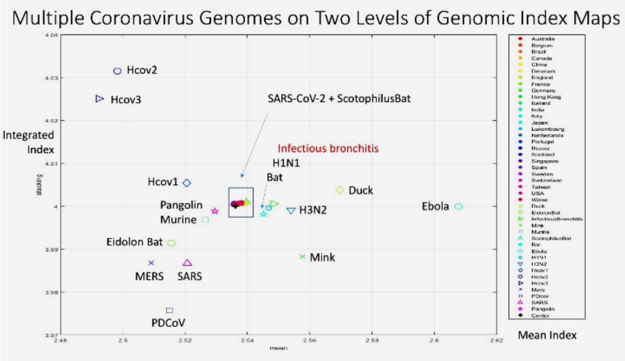

Fig. 2. Forty-Three Coronaviruses of Enlarged Figure 1(a) on Genomic Index Maps m = 16.

In Fig. 1(b), 27 genomes are located in the enlarged region. Only a Scotophilus bat genome (a blue triangle) is on the rightist side of the area. The other 26 genomes identified by various color points from different countries can be separated on most positions. Two genomes (2.5363, 4.0004) overlapped. It is interesting to see further clustering effects for larger numbers of genomes on the maps under various scaling operations.

Diagnosis and pathogenesis aspects

It is interesting to be compared with classifications of multiple coronavirus in [**10**] on similarity network analysis technologies extensive computations required. In coronavirus mutations affecting the deadliness of strains [**11**] for medical treatments of COVID-19 patients, distinguished clusters are based on phylogenetic trees identified. In both applications, genomic index maps can provide simple and clear relationships among a set of unique genomic indexes to provide general configurations to determine the intrinsic relationships under complicated similarities.

It is a natural approach to use 2D or higher dimensional maps to make refined classifications for coronavirus genomes from multiple regions in suitable classes.

Since multiple coronavirus datasets and 2D maps are involved in both CpG deficiency [**16**] and proposed genomic index maps, it is helpful to make a brief comparison on both schemes. Different from a CG proportion associated with one global invariant of a genome, both the mean index and integrated index are two global variants associated with one segmented length parameter m to provide a series of a global invariant family. In relation to two 2D maps, CG & CpG maps have a maximal size limitation. However, all possible 2D genomic index maps have the capacity to perform infinite scaling operations on any selected region. This provides super revolutionary ratios to visual possible variations of viral evolutions in refined details.

From both maps, CpG 2D maps indicate SARS-CoV-2 genomes located on the lowest part of the whole map, and (MI, II) index maps have shown SARS-CoV-2 located in the center position relevant to all selected coronavirus genomes. Considering a family of global invariants associated with genomic index maps, extremely richer variations of combinations, mutations, and evolutions could be visualized on genomic index maps to explore the original and intermediate hosts for SARS-CoV-2 genomes in COVID-19 environments.

Benefits and drawbacks on genomic index maps

In relation to Diagnosis and Pathogenesis applications, significant benefits can be observed to use genomic index maps much easier and clearer than Phylogenetic trees to show hierarchical organization among multiple genomes as selected invariant projections. From a computational viewpoint, all suitable genomic indexes are generated from relevant whole sequences without uncertain elements in the current GISAID databank with only 1/6 genomes to meet requirements, which could be a drawback for genomic indexes compared with some phylogenetic programs to organize phylogenetic information to use only partial key sequences.

Conclusion

Genomic index maps are useful to determine complicated spatial relationships among multiple genomes. The origin and variations of SARS-CoV-2 genomes have attracted attention in scientific research and medical practice, and it is important to provide hierarchical tools for precise analysis on larger datasets of genomes to analyze SARS-CoV-2 genomes and other probable intermediate hosts naturally. Refined classifications and clustering information provide precise medical treatments for various patients with infectivity by different strains of SARS-CoV-2 genomes on mutation variations of hosts worldwide. From an evolutionary viewpoint, the genomic index of a Scotophilus Bat genome [**23**] has a closer position to SARS-CoV-2 genomes than both Pangolin [**24**] and Bat SL-CovZC45 [**6**] genomes. Further refined explorations are required. It may provide a useful tool for refined visualization to explore original and intermediate hosts of SARS-CoV-2 genomes in genomic index maps for COVID-19.

Acknowledgements

The authors would like to thank NCBI, GISAID, CNGBdb, Nextstrain and Zhigang Zhang for providing invaluable information on the newest dataset collections of SARS-CoV-2 and other coronavirus genomes to support this project working smoothly. This work was supported by the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018ZI002).

Conflicts of interest

There are no conflicts to declare.

Keywords

Viral evolution, genomic index, visual maps, projection, information entropy, (mean, integrated) index.

Received: 21 June 2020
Revised: 02 July 2020
Accepted: 12 August 2020

Supporting information

Supporting information is available online at the journal website.

References

1. Andersen, K.G.; Rambaut, A.; Lipkin, W.I.; et. al. *Nat Med*, **2020**, *26*, 450.
2. Zhang, T.; Wu, Q.; Zhang, Z.; *Curr. Biol*., **2020**, *30*, 1346.
3. Laiolo, P.; Pato, J.; Jimanez-Alfaro, B.; et. al. *Nat. Commun.*, **2020**, *11*, 882.
4. He, W.; Xiang, J.; He, W.; et. al. *Molecular Biology and Evolution*, **2020**, msaa117.
5. Maganga, G.D.; Pinto, A.; Mombo, I.M.; et. al. *Sci. Rep.*, **2020**, *10*, 7314.
6. Hu, D.; Zhu, C.; Ai, L.; et. al. *Emerg Microbes Infect*, **2018**, *7*, 154.
7. Zhou, P.; Yang, X.L.; Wang, X.G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.R.; Zhu, Y.; Li, B.; Huang, C.L.; et. al. *Nature*, **2020**, *579*, 270.
8. Carro, L.; Nouioui, I.; Sangal, V.; et. al. *Sci. Rep.*, **2018**, *8*, 525.
9. Guindon, S.; Gascuel, O.; *Syst Biol.*, **2003**, *52*, 696.
10. Li, C.; Yang, Y.; Ren, L.; *Infect Genet Evol.*, **2020**, *82*, 104285.
11. Yao, H.P.; Lu, X.Y.; Li, L.J.; medRxiv, **2020**, *DOI: 10.1101/2020.04.14.20060160*.
12. Masters, Paul; *Advances in Virus Research*, **1999**, *53*, 245.
13. Aguilar-Rodriguez, J.; Peel, L.; Stella, M.; Wagner, A.; Payne, J.L.; *Evolution*, **2018**, *72*, 1242.
14. Bank, C.; Hietpas, R.T.; Jensen, J.D.; Bolon, D.N.; *Mol Biol Evol.* **2015**, *32*, 229.
15. Boyle, EA.; Li, YI.; Pritchard, JK.; *Cell*, **2017**, *169*, 1177.
16. Xia, X.; *Mol. Biol. Evol*., **2020**, *37*, 2699.
17. Ryan, T.; Koehler, Nicolas Peyret, *Bioinformatics*, **2005**, *21*, 3333.
18. Chandak, S.; Tatwawadi, K.; Weissman, T.; *Bioinformatics*, **2018**, *34*, 558.
19. Ochoa; Hernaez, M.; Weissman, T.; *Bioinformatics*, **2015**, 31, 626.
20. Zheng, J.; Zheng, C.; Biometrics and Knowledge Management Information Systems, Chapter 11: Variant Construction from Theoretical Foundation to Applications, Springer Nature **2019**, 193-202
21. Zheng, J.; Zhu, M.; Qiao, M.; Zhou, Y.; Visualizations of SARS-CoV-2 Genomes on Genomic Index Maps, submitted to Electric Current Neurology, in Review Process. **2020**, DOI: 10.21203/rs.3.rs-65159/v1.
22. Zhu, M.; Zheng, J.; Cluster Analysis of Visual Differences on Pairs of SARS-CoV-2 Genomes, submitted to Electric Current Neurology, in Review Process, **2020**.
23. Dijkman, D.; Hock, L.; *Journal of the Formosan Medical Association*, **2009**, *108*, 270.
24. Zheng, J.; Zhou, Y.; Zhu, M.; Qiao, M.; Zhang, Z.; Spread of SARS-CoV-2genomes on genomic index maps of hierarchy. Submitted to Nature Scientific Report in Review Process, **2020** preprint https://www.researchsquare.com/article/rs-31883/v1

Authors biography

Dr. Jeffrey Zheng, received ME and PhD degrees from USTC in 1981 and Monash in 1994. He has been a professor at the School of Software, Yunnan University, China, since 2004. He is a member of IEEE, SPIE and a FIAAM member of IAAM. His research has focused on variant construction. He has received numerous awards, including Scientific Creative Excellent (2007/2012); 2012 Higher Levels of Overseas Scholar project in Yunnan.

Miss Minghan Zhu is graduated from Heilongjiang University 2018. She is a postgraduate in School of Software Yunnan University supervised by Professor Jeffrey Zheng to focus attention on genomic index schemes of multiple information entropy mechanism. She has published several papers on Photon Optics, Statistics, Quantum Applications and Computer Science in various Conferences and Journals.