Different genomic representations of novel pathogens base on signal processing algorithms: COVID-19 case study

Rabeb Touati
University of Tunis El Manar
Mohamed Touati
University of Tunis El Manar, National School of Engineers of Tunis
Faouzi Benzarti
University of Tunis El Manar, National School of Engineers of Tunis
Vishal Kumar
Bipin Tripathi Kumaon Institute of Technology
Maher Kharrat
University of Tunis El Manar
Ahmed A. Elngar (elngar_7@yahoo.co.uk)
Beni-Suef University

Research Article

Keywords: Coronavirus, COVID-19, image recognition, SARS-Cov-2, RATG13, Pangolin CoV GX-P2V, Pangolin CoV MP789, CoVZXC21, CoVZC45.

Posted Date: June 23rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1743456/v1

Abstract

Coronaviruses are a type of frequent RNA virus. They are responsible for digestive and respiratory infections in animal and human genomes. A coronavirus renamed COVID-19 appeared and spread in the world which makes it declared in March 2020 a pandemic by the World Health Organization. SARS-CoV-2 genome, responsible for COVID-19 virus, has a size equal to 29,903 nucleotides, and its genetic makeup is composed of 11 functional open reading frames (ORFs). This paper proposes comparison algorithms to find SARS-Cov2 origin using digital genomic signatures. As a result, the five closest genomes related to SARS-CoV-2 are RaTG13, Pangolin-cov-PCoV_GX-P2V, Pangolin CoV MP789, bat-SL-CoVZC45 and bat-SL-CoVZXC21.

I. Introduction

Coronaviruses are responsible for digestive and respiratory infections. These RNA viruses (single-stranded positive-sense) are known to contain two of the biggest viral genomes in size (27–32 kbp) [1–10]. The coronavirus spike proteins specificity of the virus's Coronaviridae family. They are multifunctional molecular machines that mediate coronavirus entry into host cells. Now Coronaviridae family contains 4 genera (ICTV). The two genera Alpha-CoV and Beta-CoV can infect mammalian hosts, while the Gamma-CoV and the recently defined Delta-CoV principally infect avian species [6–10].

The phylogenetic study results have proved a complex history of the coronavirus's evolution that is thought to have olden origins that can in time evolution lead to cross-species infection [7–12]. The diversity of sources of the coronaviruses is presented in the ability of mutation that permits transmission of the infection from bats and birds to other species like mammals and humans [4, 7–13].

A high mutation rates, during the replication of RNA-dependent RNA polymerases (RdRP, RDR) (or renamed RNA replicase) which play a major role not only in viral replication but also in the viral RNA genetic evolution [7–16].

Coronaviruses used a special mechanism "template switching" that contributes, among their viral genomes, to high rates of homologous RNA-recombination [9, 16–20]. In addition, the largest size of these types of genomes (coronaviruses) is thought to be capable to admit gene mutations [6–8]. Today, all these factors can contribute to the variety and the plasticity of coronavirus species. Human coronaviruses are highly pathogenic; coronavirus causes SARS-CoV and coronavirus causes a MERS-CoV to belong respectively to Beta-CoV lineage B "sub-genus Sarbecovirus" and Beta-CoV lineage C "sub-genus Merbecovirus" [9, 21–23].

In March 2020, a novel worldwide pandemic (caused by the SARS-CoV-2 genome) was declared by WHO (https://covid19.who.int/). This novel disease was later identified as coronavirus called "COVID-19". COVID-19 has been spread around the world, ~ 223 countries, causing a huge health crisis [24–31]. SARS-CoV-2, like another family of CoVs, is an enveloped, positive-sense, single-stranded RNA virus. SARS-CoV-2 genome belongs to the Coronaviridae family, the Orthocoronavirinae sub-family, and the Beta-CoV genus [24]. The viral genome of 29,903 nucleotides, approximately, contains 5'and 3' untranslated regions and 11 ORFs encoding 11 proteins: "orf1ab", "S", "ORF3a", "E", "M", "ORF6", "ORF7a", "ORF7b", "ORF8", "N", and" ORF10".

Recent researches indicate a high correlation between the SARS-CoV-2 genome and the two coronaviruses genomes RaTG13 and MP789 Pangolin followed by 2 genomes: CoVZC45 and CoVZXC21 [31–33]. Till now, scientists research how SARS-CoV-2 spread to people. For this, novel comparative research employing the whole genome that searches the strain the COVID-19 confirms that belongs to lineage B (Sarbecovirus) of Beta-CoV genera. A recent Comparative study proposes that the SARS-CoV-2 genome is recombination between 2 viruses (Bat and Pangolin) [31].

In this study, we use the combination of special genomic representations (bioinformatics) and signal processing tools to have different graphical representations with the aim to understand the intragenic variations between different coronavirus genomes and to search the origin of the SARS-CoV-2 genome. This graphical representation makes it easy to identify the similarity between the human COVID-19 coronavirus and the other coronavirus families: Beta-CoV, Alpha-CoV, Gamma-CoV, and Delta-CoV. After identifying the nucleotide variation between genomes, we can classify the order of the similarities between COVID-19 and all relevant viral genomes (25 genomes) using 1D and 2D genomic signatures without any prior biological knowledge. This classification order is based on different coding techniques, including CGR, applied to the genomic sequence that corresponds to a DNA sequence.

Ii. Material And Methods

Graphical signature representation is a very important step to know the variation between different genomic sequences [31, 34–40]. The major advantage of this step is being unworthy of any need for any previously biological knowledge. The only needed here is a biological database (DNA sequences) and knowledge of the two bioinformatics and signal processing tools. We start by creating our genomics database which contains 25 investigated virus sequences. After that, we applied to each DNA sequence different coding techniques. Then we apply an analysis technique including genomic signals to see the correlation between the genomes.

1. Genomic database:

In this step, all the investigated coronavirus genomes were downloaded from the public NCBI database (http://www.ncbi.nlm.nih.gov/Genbank/). Our constructive database contains 25 genomes including the COVID-19 genome (Fig. 1).

2. Genomic Image Representations:

Obtain numerical DNA representations using different methods were used as an important step that specializes in DNA sequences if it contains repetition patterns [34–40]. To obtain different numerical representations (2-D) of DNA sequences we can use different methods: CGR image, or analysis technique applied to DNA signals. To obtain a DNA signal we can use different coding techniques: the binary [31], the EIIP mapping [41], the structural bending trinucleotide (PNUC) [42], the Frequency Chaos Game Signal (FCGS) [34–40], and so on.

2.1 CGR representation

The CGR image represents one DNA sequence unambiguously along with other sequences that reveal both local and global patterns hidden in it [31, 34–45]. This simple representation of DNA sequence is derived from chaos theory which was proposed, in 1990, by Jeffrey and it was considered as a mapping method of genome sequences [34, 43–45].

To construct a 2-D image for a DNA sequence, an iterative mappingmethod assigns to each nucleotide a 2-D coordinate (X, Y) in 2-D dimensional space [43–45]. This constructed image containthe distribution of the dotson captured in a form of 0 (empty coordinate) to 1 (dot Coordinate) square matrix. M nucleotides (N_i=N_1, N_2, .., N_M) is represented into a square by a point CGR_n, which N can be A or T or C or G. This point N_i nucleotide position in 2-D dimensional spice (X, Y) is repeatedly placed halfway between the previous plotted point N_{i-1} and the segment joining the vertex corresponding to the read letter N_i [34]. The following formula presents the prolific iterative CGR function.

$$\text{Squarewitheachcorner}\begin{cases} X_A = (0,0) \\ X_T = (1,0) \\ X_C = (0,1) \\ X_G = (1,1) \\ P_0 = (0.5, 0.5) \end{cases} \quad \text{; nucleotide position (N) in 2-D dimensional space}$$

$$\mathrm{N_i} = \begin{cases} \mathrm{N_i} = (\mathrm{X_i}, \mathrm{Y_i}) \\ \mathrm{X_i} = 0.5 * (\mathrm{x_{Nucleotide}} + \mathrm{x_{i-1}}); \\ \mathrm{Y_i} = 0.5 * (\mathrm{y_{Nucleotide}} + \mathrm{y_{i-1}}) \\ i = 1, 2, .. M \end{cases}$$

Figure 1 presents the steps (5 steps) of applying CGR representation to DNA sequence, which the obtained result is 5 points.

To compare the similarities between more of two DNA sequences, we propose to calculate center point (centroid) of 4^k squares of the CGR image for each sequence. Each centroid corresponds to sub-image after partitioning the CGR image into (k/4) × (k/4) equal sub-images. After that, for each sub-region, we compare the centroid value of each sub-region to knowing the must similar sequences. Each partition of this sub-region contains a local information and if we divide the image to k image and calculate the center we can find the CGR centroid correspond to each nucleotide (A,T,C or G). Here two-point are within the same quadrant correspond to a succession of nucleotides in the sequence with the same last mononucleotide. In addition when these points are within the same sub-quadrant, the DNA sequences have the same last dinucleotides; and so on. The coordinate of the centroid point corresponds to local information of the sub-region and can differentiate the sequence and can be used to knowing the degree of similarity between DNA sequences. Then, for each sub-region, we calculate all pairs of distances between the Covid-19 centroids and the others centroids sequences. The distance between them can indicate their similarities degree. The following flowchart presents the GGR Centroid steps (Fig. 3).

2.2 FCGR representation

CGR image contains the specific genomic signature of a given DNA sequence. After dividing this image into 4^k squares, we can obtain a FCGR representation that presents the global information of the DNA sequence. Each sub-square is associated to a sub-pattern and has a side of $\left(\frac{1}{2}\right)^k$. A visible pattern in the $FCGR_k$ corresponds to some specific pattern of a DNA sequence. The FCGR image order 1 to order 4 is presented in Fig. 4.

2.3 Time-frequencies analysis technique:

The time-frequency (T-F) analysis techniques are vital step to visualize hidden information's in DNA or RNA or proteins signals [31, 35–40]. First to obtain a signal from a genomic sequence we can use diverse coding techniques: the electron-ion interaction pseudo-potential (EIIP) mapping [41], the binary [31], the structural bending trinucleotide (PNUC) [42], the Frequency Chaos Game Signal (FCGS) [34–40], and so on.

After that, several analysis techniques can be applied to the obtained signal: Fourier Transform (FT), Wavelet Transform (WT), S transform, and so on. In this paper we have used the EIIP as a coding technique and for analysis technique we have used, the Smoothed Discrete Fourier Transform (SDFT) [31, 36] and the Continuous Wavelet Transform (CWT) [31, 35–37]. For this, a genomic sequence (nucleotide and protein sequences) is converted into a 1-D signals before processing. The investigated genomic sequences are extracted from the public NCBI platform.

Table 1
EIIP coding technique for transformation of the genomic sequence into a signal.

Amino Acid	Single Letter Symbol	EIIP amino-acid value	Amino Acid	Single Letter Symbol	EIIP amino-acid value
Ala	A	0,0373	Leu	L	0
Arg	R	0,0959	Lys	K	0,0371
Asn	N	0,0036	Met	M	0,0823
Asp	D	0,1263	Phe	F	0,0946
Cys	C	0,0829	Pro	P	0,0198
Gln	Q	0,0761	Ser	S	0,0829
Glu	E	0,0058	Thr	T	0,0941
Gly	G	0,005	Trp	W	0,0548
His	H	0,0242	Tyr	Y	0,0516
Ile	I	0	Var	V	0,0057

Nucleotide	EIIP
A	0.126
G	0.0806
C	0.1340
T	0.1335

After transforming these genomic sequences processing, the SDFT transform, which is based on Discrete Fourier Transform, have been applied to genomic signal [31, 36]. Figure 5 and Eq. 1 illustrate the SDFT steps applied to DNA numerical sequence in the aim to obtain a time-frequencies representation corresponding to a DNA sequence.

$$\begin{cases} R = 512, \mathrm{frameslength}, & \Delta \mathrm{r} = 256; \mathrm{shiftindex}, \\ N = 64, \mathrm{subframeslength}, & \Delta \mathrm{n} = 32 \end{cases} ; \mathrm{windowtype} = \mathrm{Blackmanwindow} \mathrm{Eq.1}$$

The Continuous Wavelet Transform (CWT), with 64 scales and the parameter $w_0 \sim 5.5$, was applied to the genomic signal in the aim to obtain a DNA image (scalogram) [31, 35, 37]. After applying a CWT analysis technique to a DNA signal we obtain a scalogram with their correspond matrix. After that, we explore the obtained time-frequency information located in DNA image by calculation of the scale-energy of scalogram in the goal to obtain a vector (1-D spectrum) with size equal to 64 that contains the energy of the DNA scalogram (wavelet matrix) by scale [37].

2.4 Recombination analysis

To more detect if the similarity exist between our investigated sequences we have used Clustal X program as an aligner after that we have analyze them using Simplot program [46] with a default settings (window size equal to 200, replicate used equal to100, a step size equal to 20, tree model is “Neighbor Joining”, distance model is ”Kimura”, gap stripping is “on”). To see shorter alignments, we have used Blast sequences tool with an Expect threshold (E-value) of 10, according to the stochastic model of Karlin and Altschul (1990) [47–48].

Iii. Results And Discussion

Using signal processing tools become an important step to characterize genomic sequences. In this paper, we convert each genomic sequence (25 species) in numerical form using CGR image (2-D). Visually, we can confirm with these CGR signatures (Fig. 6) the similarities between the SARS-CoV-2 genome and five coronavirus genomes: BetaCov-RaTG13 (B), Pangolin-cov-PcoV_GX-P2V (X), Pangolin_CoV-MP789 (V), bat_SL_CoVZC45 (C) and bat_SL_CoVZXC21 (D). BetaCov-RaTG13 (B), Pangolin-cov-PcoV_GX-P2V (X) and Pangolin_CoV_MP789 (V) genomes are the nearest relatives to SARS-CoV2 genome discovered so far. Sars-cov-2 genome is closer to the Beta_Cov-RaTG13 genome.

The Fig. 3 shows the plot of CGR representation of the closest sequences relatives to human SARS-CoV-2 genome. Visually, the CGR representations show the existence of similarities between sars-cov-2 genome and

5 species (B, V, X, C, D). These CGR representation confirm the obtained results of CGR-centroid method.

In another hand, we can visualize the DNA sequence in 1D (specter) by applying to the EIIP signals two analysis techniques ; the energy formula [39] to the scalograms of Continuous Wavelet Transform (CWT) and the SDFT. After computing the correlation value between 1-D specter of SARS-Cov-2 genome and our investigated spectra of DNA sequences we can class our investigated sequences to four groups; the first contains A,B,C,D,V and X viruses ; the second contains I, J, K, L, and U viruses ; the third contains S, Y, Z, R, Q, and H viruses ;the four contains G and O viruses. Figure 8 shows the spectra superposition of our investigated sequences and the greatest similar specter to SARS-Cov-2 genome. The given 1-D spectra reflect the existence of similarities between these sequences (B, C, D, V, and X) an SARS-CoV-2 genome. These obtained result confirms the obtained results of CGR representation.

For more investigation, we use visualize the degree of similarity and to confirm the possibility of the recombination between RATG 13 and pangolin MP789 sequences, we have used Simplot analysis. Figure 9 shows the greet similarity between SARS-Cov-2 genome and the two coronavirus genomes RATG13 and pangolin. However, we can see that the Yak Beta-CoV (U) seems to be different contrary to what was said in the article [49].

Iv Conclusion

In our work, we proposed original methods of genomic processing and identification to achieve comparison between various genomes of coronaviruses in order to check the similarities that can exist between the human Covid-19 and other viruses of the same family, affecting other species. Indeed such investigation is necessary to know the origin and evolution pattern of the SARS-Cov-2 genome. The obtained results with different DNA representations methods were compared to each other and thus results were confirmed. Furthermore, they were also compared to results obtained by conventional methods used by biologists such as Blast comparison and Simplot analysis to identify possible recombination event. Accordingly, our results were confirmed which demonstrates the exactness and robustness of our algorithms. In addition, for the CGR-Centroid point's, the coordinate value in a 2-dimensional space (x, y) of each point can constitute an input for the classification system to classify different genomics sequences. A new proposed algorithms that are based on nucleotides frequencies can be used in the goal to classify and identify other DNA sequences types. We can classify DNA sequences by calculate the correlation spectra between these sequences. The spectra correlation results are the existing of groups and each group presents the more similar viruses ; (1) we have A, B, C, D, V and X viruses ; (2) we have I, J, K, L, and U viruses ; (3) S, Y, Z, R, Q, and H viruses ; (4) G and O viruses.

These phylogenetic trees results, using classification based on numerical DNA sequences, reflect the performance of our algorithms. Our methods can be used to solve biological problems.

Abbreviations

ICTV	International Committee on Taxonomy of Viruses
CoV:	Coronaviruse
CoVs :	Coronaviruses
Alphacoronavirus	Alpha-CoV
Betacoronavirus	Beta-CoV
Gammacoronavirus	Gamma-CoV
Deltacoronavirus	Delta-CoV
COVID-19	Coronavirus disease 2019
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SARS-CoV	Severe Acute Respiratory Syndrome
MERS-CoV	Middle East respiratory syndrome
ORFs	Open Reading Frames
S	Spike
ORFs	Open Reading Frames
CWT	Continuous Wavelet Transform
EIIP	Electron Ion Interaction Potential
PCV10	Streptococcus pneumonae
NCBI	National Center for Biotechnology Information
WHO	World Health Organization
CGR	Chaos Game representation
BLAST	Basic Local Alignment Search Tool
EIIP	electron-ion interaction pseudo-potential

Declarations

Ethics approval and consent to participate

This study did not include Human participants or Patient data. Hence no ethical approval and consent to participate is required.

Consent for publication

Not applicable. This study did not include patients.

Availability of data and materials

All data generated or analyzed as part of this study are included in this published article.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the Tunisian Ministry of Higher Education and Scientific Research (Research laboratory: LR99ES10).

Authors' contributions

all authors' individual contributions, using the relevant CRediT roles:

Conceptualization: Rabeb Touati

Data curation: Mohamed Touati

Formal analysis: Faouzi Benzarti

Investigation: Rabeb Touati

Methodology: Maher Kharrat

Project administration: Ahmed A. Elngar

Resources: Mohamed Touati

Software: Rabeb Touati

Supervision: Ahmed A. Elngar

Validation: Faouzi Benzarti

Visualization: Vishal Kumar

Roles/Writing - original draft: Rabeb Touati

Writing - review & editing: Ahmed A. Elngar

Acknowledgements

The authors are thankful to the co-authors for their support to complete the manuscript. We are also thankful to the anonymous reviewers for their perceptive comments on the manuscript.

References

1. Weiss, S. R., & Navas-Martin, S. (2005). Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. Microbiol. Mol. Biol. Rev., 69(4), 635–664.
2. Perlman, S., & Netland, J. (2009). Coronaviruses post-SARS: update on replication and pathogenesis. Nature reviews microbiology, 7(6), 439–450.
3. Su, S., Wong, G., Shi, W., Liu, J., Lai, A. C., Zhou, J., … Gao, G. F. (2016). Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends in microbiology, 24(6), 490–502.
4. Cui, J., Li, F., & Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. Nature reviews Microbiology, 17(3), 181–192.
5. Schoeman, D., & Fielding, B. C. (2019). Coronavirus envelope protein: current knowledge. Virology journal, 16(1), 69.
6. King, A. M., Lefkowitz, E., Adams, M. J., & Carstens, E. B. (Eds.). (2011). Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses (Vol. 9). Elsevier.
7. Woo, P. C., Lau, S. K., Huang, Y., & Yuen, K. Y. (2009). Coronavirus diversity, phylogeny and interspecies jumping. Experimental Biology and Medicine, 234(10), 1117–1127.
8. Wertheim, J. O., Chu, D. K., Peiris, J. S., Pond, S. L. K., & Poon, L. L. (2013). A case for the ancient origin of coronaviruses. Journal of Virology, 87(12), 7039–7045.
9. Luk, H. K., Li, X., Fung, J., Lau, S. K., & Woo, P. C. (2019). Molecular epidemiology, evolution and phylogeny of SARS coronavirus. Infection, Genetics and Evolution.
10. Vijaykrishna, D., Smith, G. J., Zhang, J. X., Peiris, J. S. M., Chen, H., & Guan, Y. (2007). Evolutionary insights into the ecology of coronaviruses. Journal of virology, 81(8), 4012–4020.
11. Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., & Bi, Y. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet, 395(10224), 565–574.
12. Lau, S. K., Li, K. S., Tsang, A. K., Shek, C. T., Wang, M., Choi, G. K., … Wang, S. Y. (2012). Recent transmission of a novel alphacoronavirus, bat coronavirus HKU10, from Leschenault's rousettes to pomona leaf-nosed bats: first evidence of interspecies transmission of coronavirus between bats of different suborders. Journal of virology, 86(21), 11906–11918.
13. Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J. H., … Zhang, J. (2005). Bats are natural reservoirs of SARS-like coronaviruses. Science, 310(5748), 676–679.
14. Jenkins, G. M., Rambaut, A., Pybus, O. G., & Holmes, E. C. (2002). Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. Journal of molecular evolution, 54(2), 156–165.
15. Duffy, S., Shackelton, L. A., & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews Genetics, 9(4), 267–276.

16. Nagy, P. D., & Simon, A. E. (1997). New insights into the mechanisms of RNA recombination. Virology, 235(1), 1–9.
17. Rowe, C. L., Fleming, J. O., Nathan, M. J., Sgro, J. Y., Palmenberg, A. C., & Baker, S. C. (1997). Generation of coronavirus spike deletion variants by high-frequency recombination at regions of predicted RNA secondary structure. Journal of virology, 71(8), 6183–6190.
18. Pasternak, A. O., Spaan, W. J., & Snijder, E. J. (2006). Nidovirus transcription: how to make sense… Journal of general virology, 87(6), 1403–1421.
19. Cavanagh, D. (2005). Coronaviridae: a review of coronaviruses and toroviruses. In Coronaviruses with Special Emphasis on First Insights Concerning SARS (pp. 1–54). Birkhäuser Basel.
20. Lai, M. M. (1992). RNA recombination in animal and plant viruses. Microbiology and Molecular Biology Reviews, 56(1), 61–79.
21. Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., … Rollin, P. E. (2003). A novel coronavirus associated with severe acute respiratory syndrome. New England journal of medicine, 348(20), 1953–1966
22. Zaki, A. M., Van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D., & Fouchier, R. A. (2012). Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. New England Journal of Medicine, 367(19), 1814–1820.
23. Drosten, C., Günther, S., Preiser, W., Van Der Werf, S., Brodt, H. R., Becker, S., … Berger, A. (2003). Identification of a novel coronavirus in patients with severe acute respiratory syndrome. New England journal of medicine, 348(20), 1967–1976.
24. Coronaviridae Study Group of the International Committee on Taxonomy of V (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol.;5(4):536–44.
25. Li, G., & De Clercq, E. (2020). Therapeutic options for the 2019 novel coronavirus (2019-nCoV). Nature reviews Drug discovery, 19(3), 149–150.
26. Hui, D. S., I Azhar, E., Madani, T. A., Ntoumi, F., Kock, R., Dar, O., … Zumla, A. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. International Journal of Infectious Diseases, 91, 264–266.
27. Liu, T., Hu, J., Kang, M., Lin, L., Zhong, H., Xiao, J., … Deng, A. (2020). Transmission dynamics of 2019 novel coronavirus (2019-nCoV).
28. Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., … Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579(7798), 265–269.
29. Gralinski, L. E., & Menachery, V. D. (2020). Return of the Coronavirus: 2019-nCoV. Viruses, 12(2), 135.
30. Johns Hopkins, C. S. S. E. (2020). Coronavirus 2019-nCoV Global Cases. Center for Systems Science and Engineering, Johns Hopkins University. Retrieved, 8.
31. Touati, R., Haddad-Boubaker, S., Ferchichi, I., Messaoudi, I., Ouesleti, A. E., Triki, H., … Kharrat, M. (2020). Comparative genomic signature representations of the emerging COVID-19 coronavirus and other coronaviruses: High identity and possible recombination between Bat and Pangolin coronaviruses. Genomics, 112(6), 4189–4202.

32. Haddad-Boubaker, S., Othman, H., Touati, R., Ayouni, K., Lakhal, M., Ben Mustapha, I., … Triki, H. (2021). In silico comparative study of SARS-CoV-2 proteins and antigenic proteins in BCG, OPV, MMR and other vaccines: evidence of a possible putative protective effect. BMC bioinformatics, 22(1), 1–14.
33. Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J. J., … Shen, Y. (2020). Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature, 583(7815), 286–289.
34. H. J. Jeffrey, "Chaos game representation of gene structure," Nucleic Acids Res., vol. 18, no. 8, pp. 2163–2170, Apr. 1990
35. Touati, R., Messaoudi, I., Oueslati, A. E., & Lachiri, Z. (2019). A combined support vector machine-FCGS classification based on the wavelet transform for Helitrons recognition in C. elegans. Multimedia Tools and Applications, 78(10), 13047–13066.
36. Touati, R., Oueslati, A. E., Messaoudi, I., & Lachiri, Z. (2019). The Helitron family classification using SVM based on Fourier transform features applied on an unbalanced dataset. Medical & biological engineering & computing, 57(10), 2289–2304.
37. Touati, R., Messaoudi, I., Oueslati, A. E., & Lachiri, Z. (2019). Distinguishing between intra-genomic helitron families using time-frequency features and random forest approaches. Biomedical Signal Processing and Control, 54, 101579.
38. Touati, R., Messaoudi, I., Oueslati, A. E., Lachiri, Z., & Kharrat, M. (2020). Classification of intra-genomic helitrons based on features extracted from different orders of FCGS. Informatics in Medicine Unlocked, 18, 100271.
39. Touati, R., Tajouri, A., Mesaoudi, I., Oueslati, A. E., Lachiri, Z., & Kharrat, M. (2021). New methodology for repetitive sequences identification in human X and Y chromosomes. Biomedical signal processing and control, 64, 102207.
40. Touati, R., Messaoudi, I., Oueslati, A. E., Lachiri, Z., & Kharrat, M. (2020). New Intraclass Helitrons Classification Using DNA-Image Sequences and Machine Learning Approaches. IRBM.
41. S. Chakraborty, V. Gupta, Dwt based cancer identification using EIIP, 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT) IEEE, 2016, pp. 718–723.
42. Touati, R., Ferchichi, I., Messaoudi, I., Oueslati, A. E., Lachiri, Z., & Kharrat, M. (2020). Pre-Cursor microRNAs from Different Species classification based on features extracted from the image. Journal of Cybersecurity and Information Management (JCIM) Vol, 3(1), 5–13.
43. N. Goldman, Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences, Nucleic Acids Research (1993). 2487–2491.
44. J.S. Almeida, J.A. Carrico,A. Maretzek, P.A. Noble & M. Fletcher, Analysis of genomic sequences by Chaos Game Representation", Bioinformatics (2001) 429– 437.
45. Fiser, A., Tusnady, G. E., & Simon, I. (1994). Chaos game representation of protein structures. Journal of molecular graphics, 12(4), 302–304.
46. Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., … Ray, S. C. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. Journal of virology, 73(1), 152–160.

47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10.
48. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA. 1990;87(6):2264–8.
49. Dabravolski, S. A., & Kavalionak, Y. K. (2020). SARS-CoV-2: Structural diversity, phylogeny, and potential animal host identification of spike glycoprotein. Journal of medical virology, 92(9), 1690–1694.

Figures

Order	Species	Accession number	Length (bp)	
A	SARS-CoV-2	NC_045512	29,903	
B	Betacoronavirus RaTG13	MN996532	29,855	
C	Betacoronavirus CoVZC45	MG772933	29,802	
D	Betacoronavirus CoVZXC21	MG772934.1	29,732	
E	Alphacoronavirus	DQ811787	27,533	
F	Gammacoronavirus A116E7	FN430415	27,593	Orthocoronavirinae
G	Deltacoronavirus	KJ481931	25,406	
H	Bat Hp-betacoronavirus/	NC_025217	31,491	
I	Bovine coronavirus Mebus	BCU00735	31,032	
J	Human coronavirus OC43	KX344031	30,713	
K	Human coronavirus OC43 ATCC VR-759	AY585228	30,741	
L	Porcine hemagglutinating encephalomyelitis virus VW572	DQ011855	30,480	
M	Murine hepatitis virus JHM	AC_000192	31,526	Betacoronavirus
N	Rat coronavirus Parker	FJ938068.1	31,250	
O	Pipistrellus bat coronavirus HKU5/HK/03/2005	NC_009020	30,482	
P	Rousettus bat coronavirus HKU9/GD/005/2005	NC_009021	29,114	
Q	SARS-related Rhinolophus bat coronavirus Rf1/2004	NC_004718.3	29,751	
R	SARS-related palm civet coronavirus SZ3/2003	AY304486.1	29,741	
S	SARS-related chinese ferret badger coronavirus CFB/SZ/94/03	AY545919	29,739	
T	Tylonycteris bat coronavirus HKU4/HK/04/2005	NC_009019	30,286	
U	Yak coronavirus strain YAK/HY24/CH/2017	MH810163	31,032	
V	Pangolin coronavirus isolate MP789	MT121216	29,521	
w	Middle East respiratory Hu/Riyadh_KSA_4050_2015	KT026456.1	30,120	Betacoronavirus
X	Pangolin-cov-PCoV_GX-P2V	MT072864.1	29,795	
Y	Bat SARS-cov_WIV1	KF367457.1	30,309	
Z	SARS_coronavirus_BJ01	AY278488.2	29,725	Betacoronavirus

Figure 1

Investigated Genomics database (26 species) from NCBI platform (http://www.ncbi.nlm.nih.gov/Genbank/)

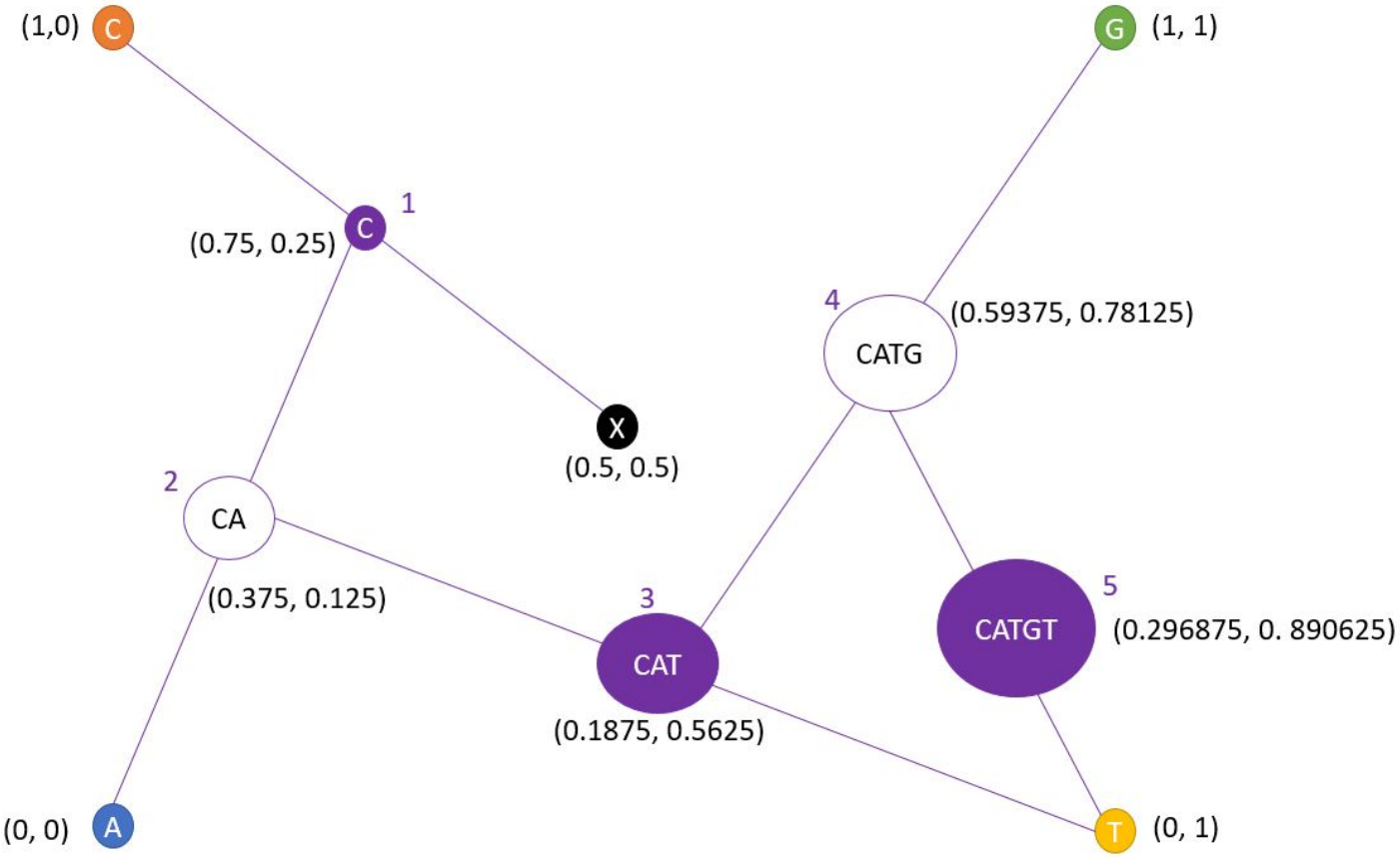

Figure 2

Illustration of the CGR process for the input ‘CATGT’ sequence

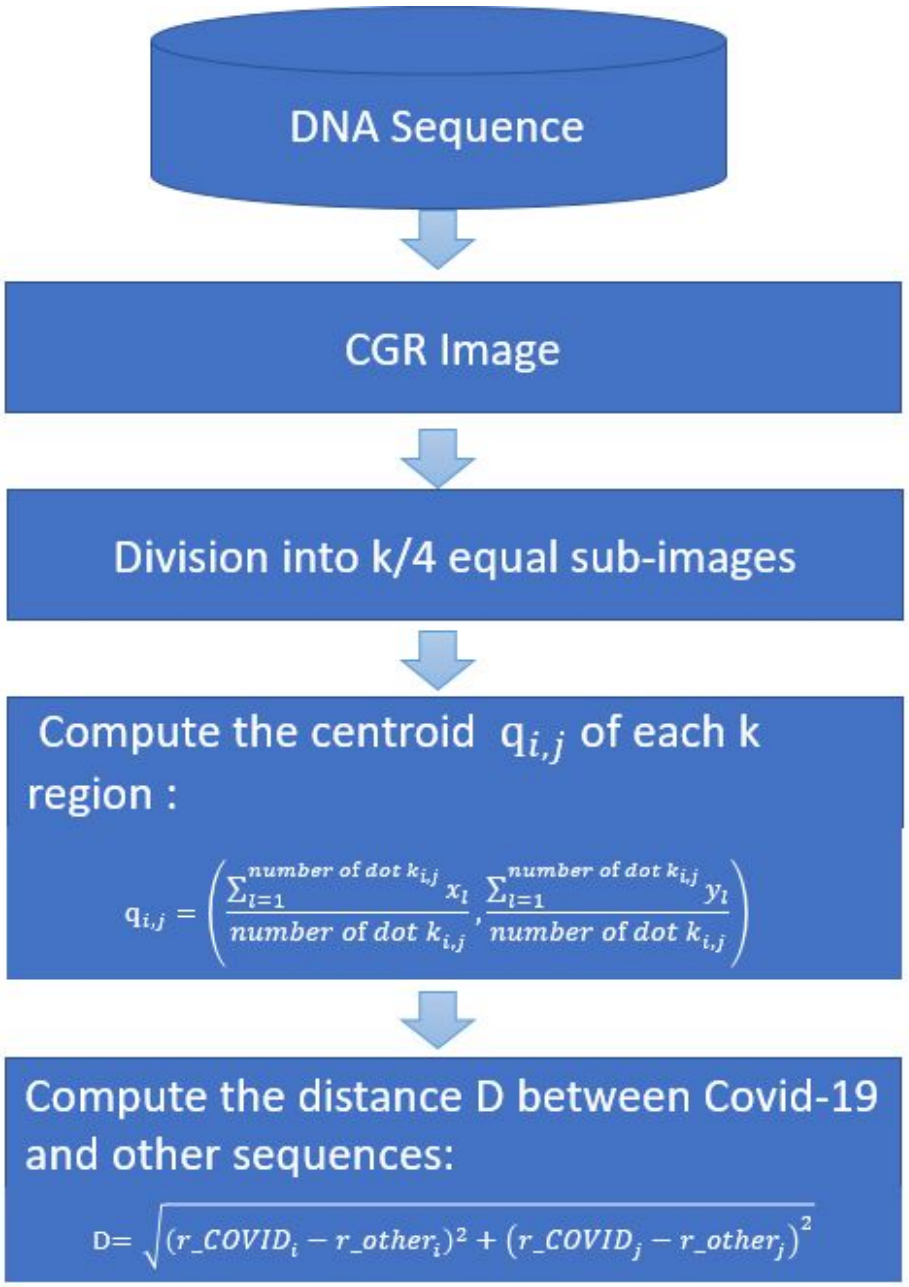

Figure 3

flowchart illustrate the GGR Centroid steps of each DNA sequence

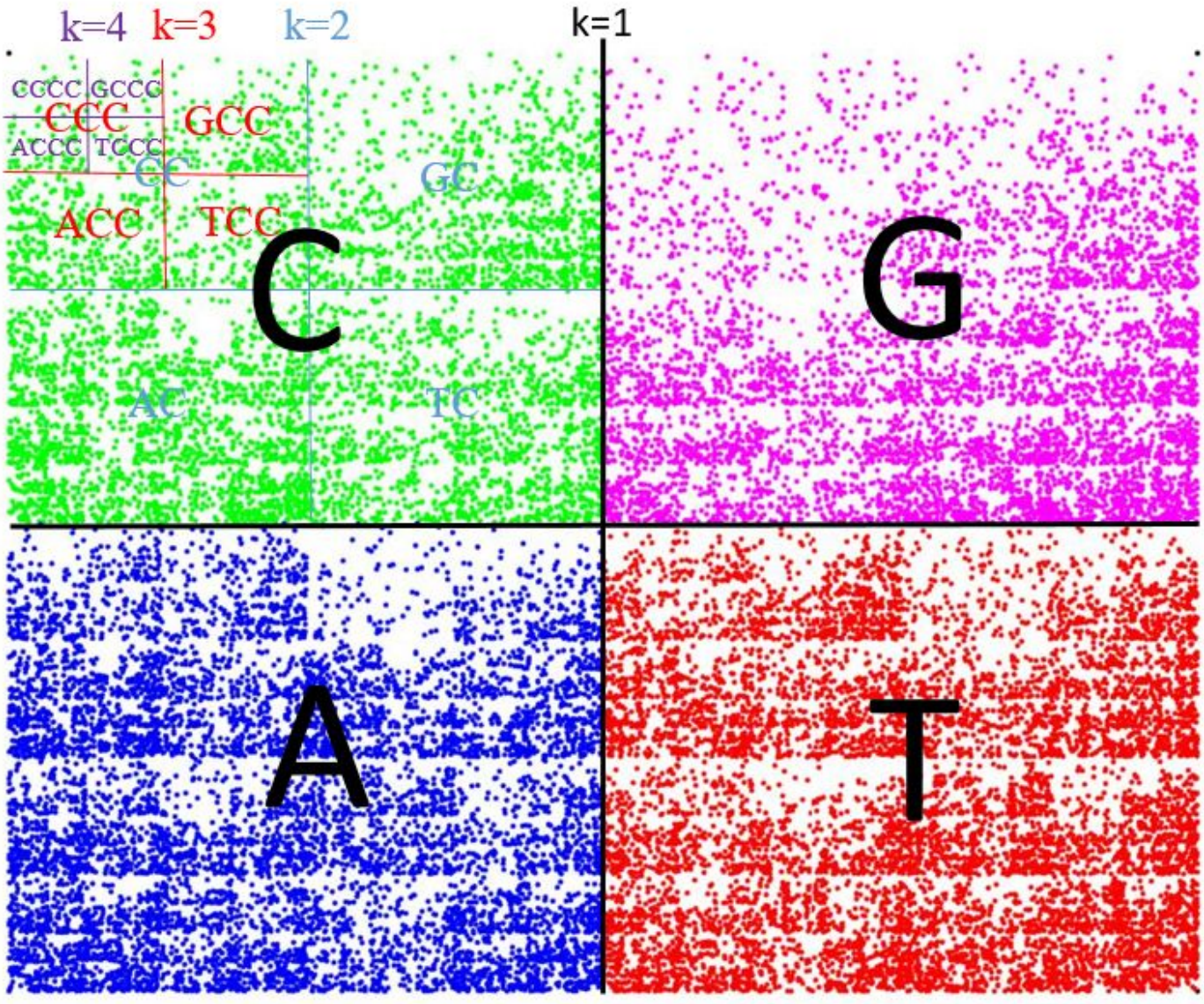

Figure 4

Frequency Chaos Game Representation (FCGR) construction on CGR image of SARS-Cov-2 genome from order 1 to order 4 (k=1, 2, 3, and 4).

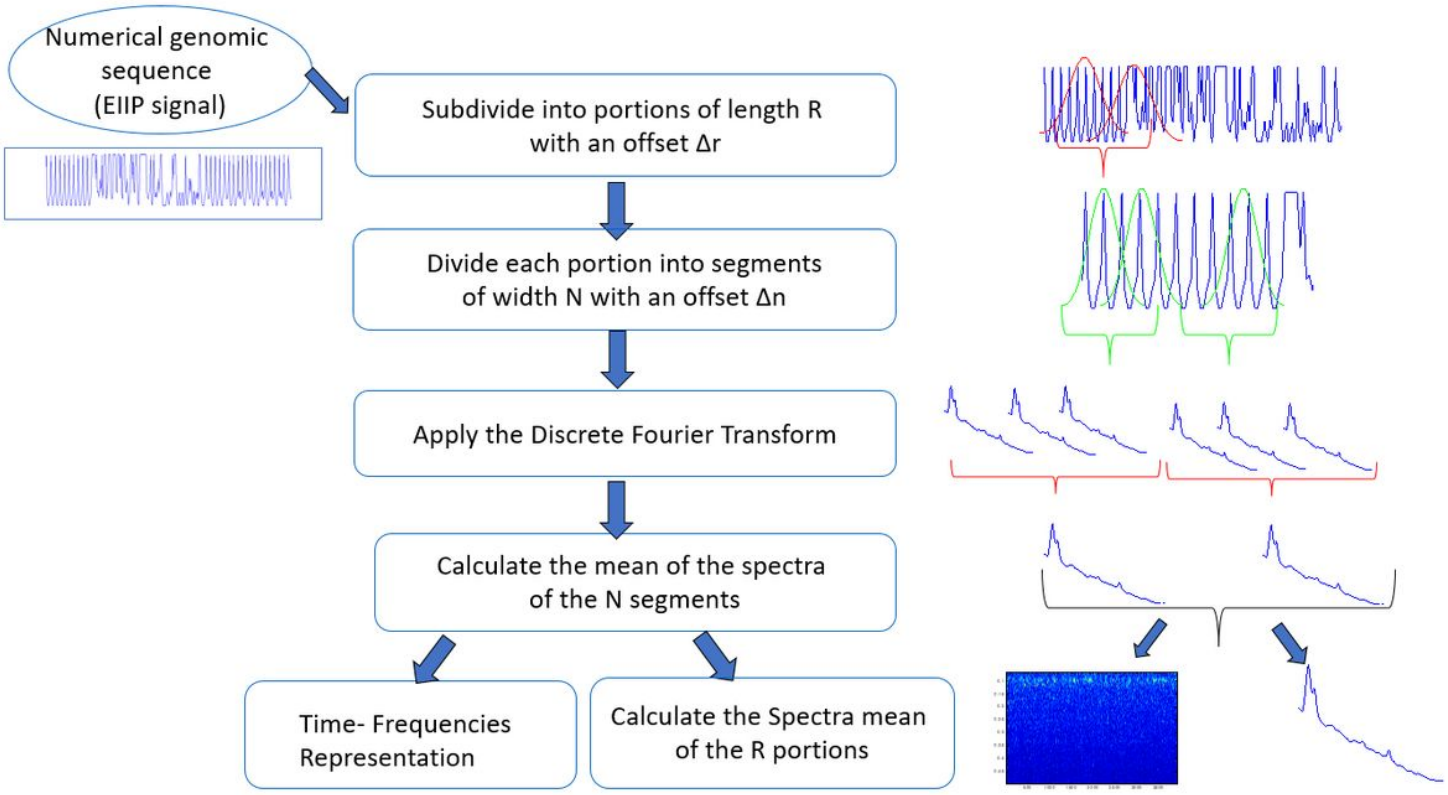

Figure 5

Flowchart of SDFT method for EIIP DNA signal

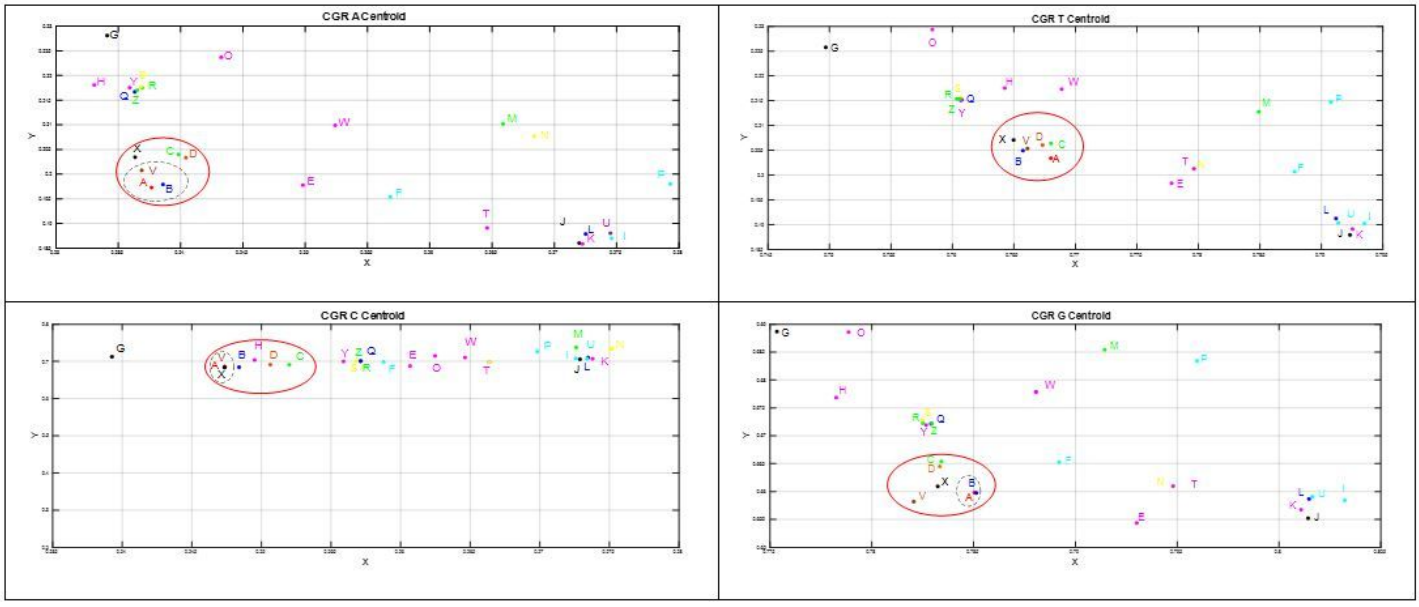

Figure 6

CGR centroids plots where the points where each subfigure shows the CGR centroid plot of A, or, T, or G or C nucleotide of our 26 investigated genomes.

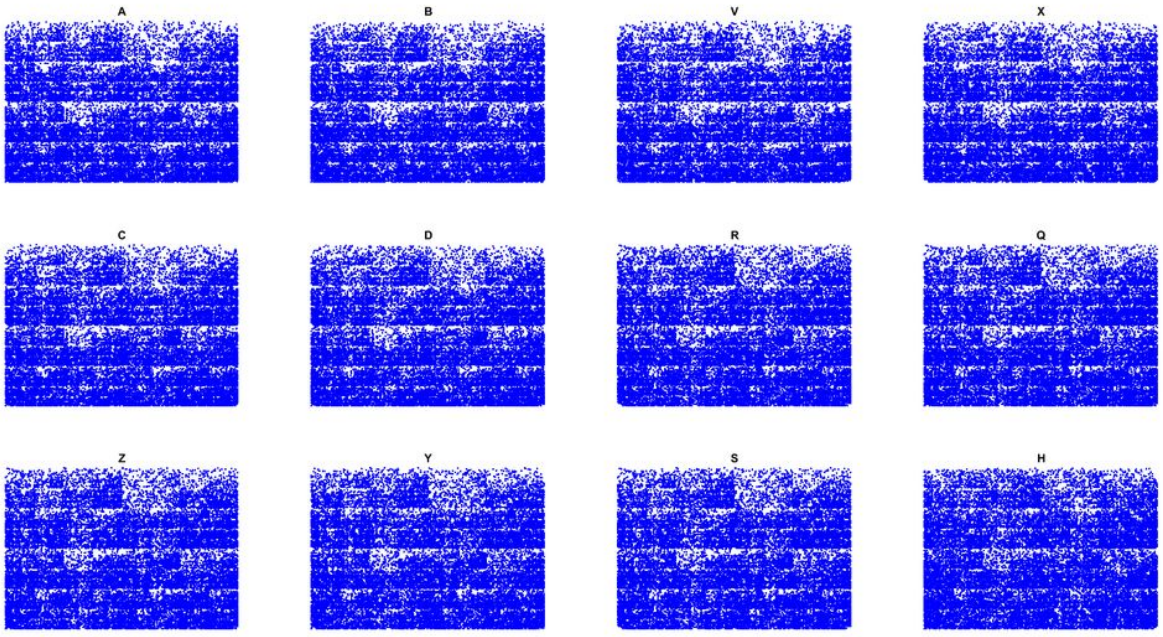

Figure 7

CGR plots of sars-cov-2 genome and their more similar species

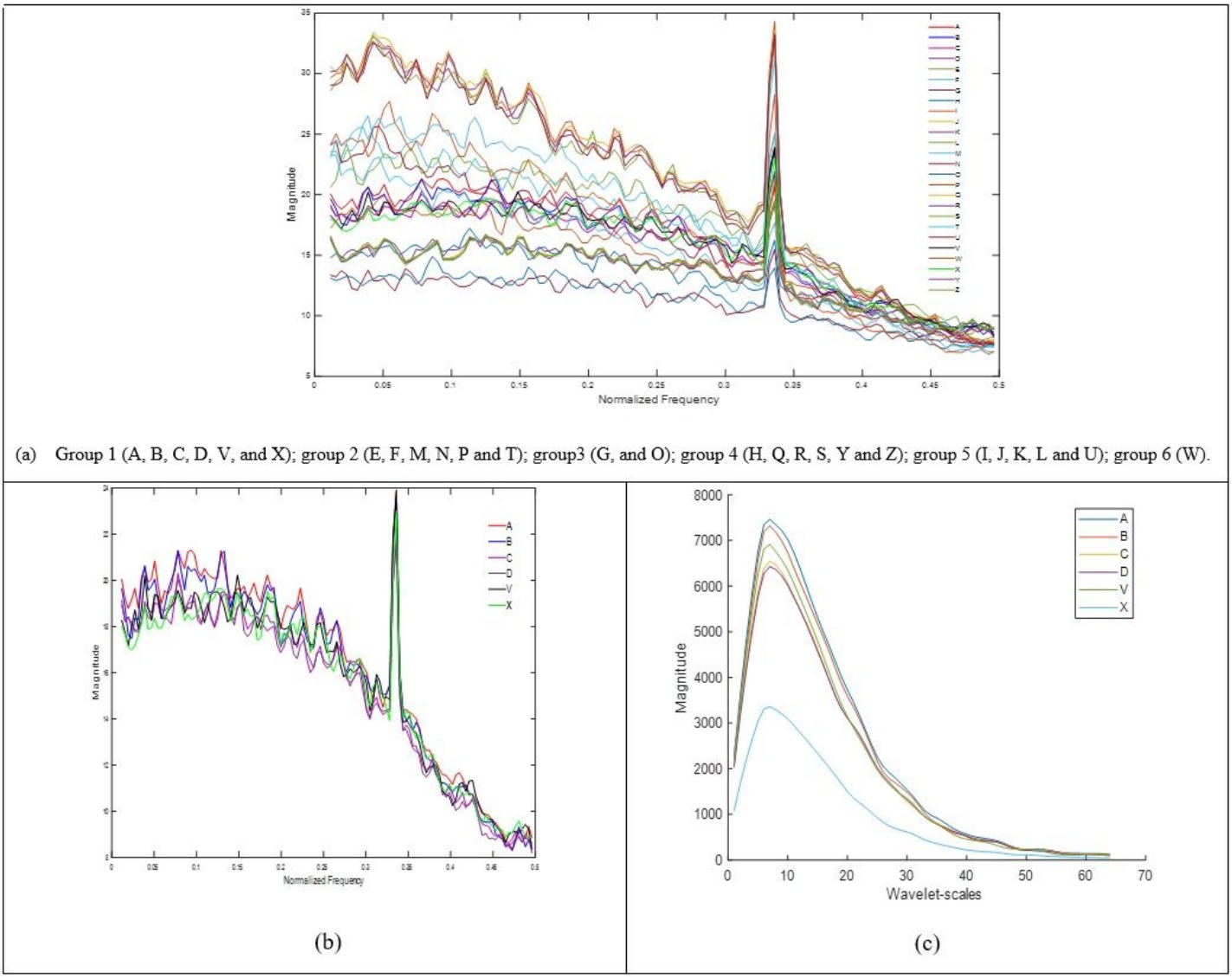

Figure 8

1-D time frequencies representations (spectra) of 26 viruses investigated in this study : (a)Viruses spectra superposition of our investigated viruses using SDFT technique (b) superposition of the must similar spectra to SARS-Cov-2 genome using SDFT technique (c) Wavelet spectra superposition of sars-cov-2 genome and their more similar species (A, B, C, D, V, AND X).

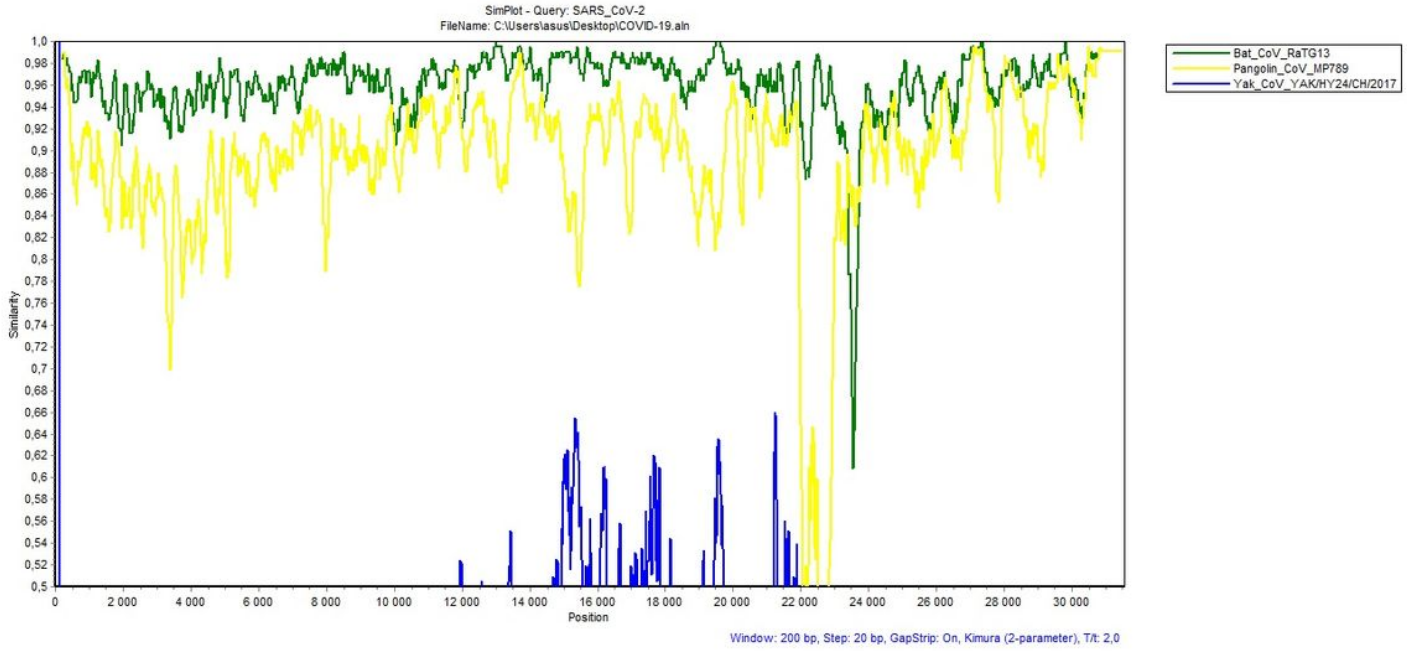

Figure 9

Simplot analysis of SARS-CoV-2 genome in comparison with 3 genomes coronavirus genomes: RATG13, Pangolin and Yak.