k-mer proximity index for phylogeny comparison of SARS-CoV-2 with other pathogens

Pratibha
Indian Institute of Technology Roorkee https://orcid.org/0000-0002-1152-6847
Cyril Shaju
Indian Institute of Technology Roorkee
Kamal (kamal@es.iitr.ac.in)
Indian Institute of Technology Roorkee https://orcid.org/0000-0003-0856-759X
Aman Gupta
Manipal University Jaipur

Research Article

Keywords: Corona Virus, COVID-19, Chaos Game Representation (CGR), Genome sequence, Epidemics,

Posted Date: June 30th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-38086/v1

Abstract

We developed a compact and computationally inexpensive method for in-silico comparison of nucleotide sequences at a macro level using subtraction-percentage plots (SP-plots) of a modified chaos game representation (CGR). Analyzing these plots, we defined the k-mer proximity index quantifying the differences between SARS-CoV-2 and other pathogens' genome sequences. We categorized 31 pathogens, on the basis of their proximity to SARS-CoV-2, in four groups to possibly plan a treatment strategy for Covid-19.

Introduction

The order of the nitrogen bases (adenine, thymine, guanine and cytosine or A, T, G and C) in a nucleotide sequence determines the genetic code of a species. One of the fundamental tasks of genomics is to compare these sequences for phylogenetic analysis. Several methods are available to compare genetic sequences, either through sequence alignment[1,2] or alignment-free approach [3,4,5,6,7]. Due to large size of genomic data, the sequence alignment approaches are not time and memory efficient and also have some shortcomings [8,9]. The alignment-free approach provide several different methodologies [10,11,12,13]. One of the approaches is based on chaos game representation (CGR) which has become popular due to its compact representation of the whole genome sequence in graphical form[14,15,16,17,18].

Here we describe a novel and simple approach to quantify the similarity/dissimilarity between two genome sequences using a modified CGR. CGR is a iterative mapping technique to convert the genome sequence of a given length into a single image based on the frequency of k-mers (the order of the nucleotides in a sequence for a word of size k) where each pixel corresponds to a specific nucleotide combination[19]. As the genome sequences of different species differ in length, a direct comparison of the species based solely on the CGR image is difficult. To make the CGR image length independent, we first convert it into a percentage CGR plot (PC-plot) by plotting, at each pixel level, the percentage of the k-mer frequencies in a genomic sequence (See Methods). Visually, the PC-plots and traditional CGR images are identical. To quantify the similarity between two species, we introduce two new concepts, a subtraction percentage CGR plot (SP-plot) and the k-mer proximity index (Pr). A SP-plot is obtained by subtracting the percentage points of respective k-mers in each sequence. The SP-plot consists of positive and negative values indicating the differences of k-mers percentage distribution between two species (See Methods). The sum of the positive differences is always be equal to the sum of negative differences. This sum is named as k-mer proximity index (Pr) which represents the degree of similarity between the genome sequences of two species (See Methods). Obviously, the value of this proximity index will increase with the degree of dissimilarity between two species. The value of this index also changes with the value of 'k' because the distribution of a specific length combination of nucleotides will change as 'k' changes.

Results

As the world is currently suffering with the deadly Covid–19 pandemic, it is important to quickly understand the interrelationship between different pathogens to plan a strategy for treatment of Covid–19 based on the available cure of existing pathogens. We compare the genome sequences of SARS-CoV–2 virus with 31 other pathogens using the visual inspection of PC-plots of individual pathogens, the SP-plots of pathogen pairs (Supplementary Figures 3–33) and quantifying the similarity through the k-mer proximity indices between each pair (Supplementary Table 1). On the basis of 4-mer proximity indices of 31 pathogens' genome sequences, we divided them in four categories, A, B, C and D, in the increasing order of their dissimilarity with SARS-CoV–2. Figure 1 (upper half of top panel) shows the PC-plot of SARS-CoV–2 genome sequence along with four other pathogens (one each from the category A, B, C and D respectively). We generated PC-plots and SP-Plots on unit squares with C, A, T and G representing corners anticlockwise starting with C at the lower left corner, C(0,0), A(1,0), T(0,1) and G(1,1). Therefore, the left halves of the squares represent pyrimidine composition and the right halves represent purine compositions. The base compositions are represented by the diagonals of the square. Although the visual difference between these images are remarkable, the SP-plots in the lower half of the top panel clearly depict the dissimilarities between the genome pairs. The SP-plots in figure 1 are obtained by subtracting the k-mer percentage points at each pixel in the respective pathogen' from those in SARS-CoV–2 image. Therefore, the red points in the SP-plots indicate a higher frequency of the corresponding k-mer in the SARS-CoV–2 genome sequence and the blue points indicate a higher frequency of respective k-mers in the pathogen which is being compared.

We calculated the SARS-Cov–2 k-mer proximity indices for several higher order oligomers (k = 4 to 9) for all 31 pathogen pairs (See Supplementary Table 1). The bottom panel of Figure 1 shows the value of 4-mer (tetra-nucleotide) proximity indices for all 31 pair of pathogens in the increasing order.

The lowest tetra-nucleotide proximity index is for Cov–2 and CoV–1 pair and the highest is for CoV–2 and Rubella pair. We divided the pathogens in four categories according to the proximity index values.

*Category A*This category has the pathogens with lowest 4-mer proximity index (less than 10) with SARS-Cov–2 and has uncanny similarity with the COVID–19 virus. The existing drugs/treatment, if available, for any of these pathogens will stand a very high chance to be successful. A suggestion to use Sofosbuvir, a potential drug for Human Corona virus (HCV) was recently reported [20] as a potential treatment for COVID–19..

*Category B*Pathogens with low 4-mer proximity index (between 10 and 20) belong to this group and any treatment available of the pathogens from this group tried with the SARS CoV–2 will stand moderate chances of success. We have already noted that for recent Remedesivir (earlier used for Hepatitis and Ebola) trials.

*Category C*Pathogens with moderate 4-mer proximity index (between 20 and 30) belong to this group and any treatment available for the pathogens belonging to this group has very little chances of success for SARS CoV–2 cure. This has been observed with a few recent trials of HIV drugs administered to Corona virus patients without much success.

*Category D*The genome sequence of the pathogens belonging to this group are quite far (4-mer proximity index more than 30) from the PC-plot of SARS CoV–2 and therefore any treatment for the pathogens in this group cannot be repeated for COVID–19. The recent failure of hydro-chloroquine (malaria) and BCG vaccine (tuberculuosis) support our hypothesis.

We plotted the variation in the value of k-mer proximity indices for 31 pathogen pairs (Figure 2). The index value increases with the increase in the size of k-mer, which means that at higher k-mer nucleotide levels, the dissimilarity between the species increases, but the relative changes in the value of k-mer proximity index for a given pathogen pair remain same throughout.

Another important aspect of this method is the time efficiency. The MATLAB code generated for this method takes only 37 seconds to compare two sequences of 30,000 points each on a core i7 laptop computer with 8 GB RAM. Our method provides a novel and compact way phylogeny comparison and quantify the similarity between two species in a time efficient way.

Methods

CGR—A genetic sequence X(k) can be considered as a string composed of A, G, C and T which represent Adenine, Guanine, Cytosine and Thymine, respectively.

$X_k \in C,A,T,G$

We consider a unit square U and name corners C_i (i = 1,2,3,4) as C, A, T and G respectively, which corresponds to the value of X(k). The initial point P(0) is the midpoint of the square. Now the second point P(1) is the midpoint between P(0) and $C_{X(1)}$. In General, P(k) is plotted as the midpoint between P(k–1) and $C_{X(k)}$[14].

After plotting the genetic sequence X in unit square U, the unit square is divided into 2^N x 2^N sub squares; each sub-square represents a unique sub-sequence of length k (k-mer).

An example for movement of points in CGR is shown with the first eight members of the data sequence (GCTTATGT) in Supplementary Figure 1. An example of addresses of the sub-squares for nucleotides, di-nucleotides (2-mer), tri-nucleotides (3-mer) and tetra-nucleotides (4-mer) is given in Supplementary Figure 2.

*PC-plots*To make these plots, the percentage of points plotted in sub-square is calculated. This percentage value represents the intensity of points in each sub-square. After plotting points by CGR and dividing the unit square into 2^k x 2^k sub squares, each sub-square is color-filled based on the calculated intensity values. Supplementary Figure 3 shows the percentage plot (Y) made for the SARS-Cov–2 and SARS-Cov–1 for k = 7. Similar plots were made for all the pathogens (See supplementary Figs 4–33).

SP-plots and k-mer proximity Index—Subtraction plot between genome1 (g1) and genome2 (g2) is plotted as

Sg1-g2 = Yg1- Yg2

For example, if percentage density values of Y_{g1} and Y_{g2} in 4x4 matrices corresponding to di-nucleotides (2-mer) are

$$Y_{g1} = \begin{bmatrix} 5 & 0 & 20 & 10 \\ 0 & 5 & 10 & 10 \\ 8 & 12 & 0 & 0 \\ 0 & 0 & 12 & 8 \end{bmatrix}, \quad \text{and} \quad Y_{g2} = \begin{bmatrix} 1 & 0 & 35 & 25 \\ 0 & 1 & 15 & 15 \\ 1 & 1 & 4 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\text{then} \quad S_{g1-g2} = \begin{bmatrix} 4 & 0 & -15 & -15 \\ 0 & 4 & -5 & -5 \\ 7 & 11 & -4 & -2 \\ 0 & 0 & 12 & 8 \end{bmatrix}.$$

From the subtraction plot S, the sum of all the positive numbers (also the sum of modulus of negative numbers) is a measure of similarity or dissimilarity between two genetic sequences.

$$Pr = \sum_{j=1}^{2^k} \sum_{i=1}^{2^k} Y_{ji} \ , where \ Y_{ji} \geq 0$$

In the above example, the 2-mer proximity index (*Pr)* is calculated by adding all positive (4+4+7+11+12+8 = 46) or all negative (15+15+5+5+4+2 = 46) values.

Declarations

Acknowledgements

Pratibha acknowledges the funding support from Science and Engineering Research Board, India.

Contributions

All authors contributed equally to the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The underlying dataset consists of DNA/cDNA of 32 genome sequences which was obtained from the NCBI genome database (http://www.ncbi.nlm.nih.gov/),

References

1. Zhi Qi, Sy Redding, Ja Yil Lee, Bryan Gibb, YoungHo Kwon, Hengyao Niu, William A. Gaines, Patrick Sung, Eric C. Greene, DNA Sequence Alignment by Microhomology Sampling during Homologous Recombination, *Cell* **160**, 856-869, https://doi.org/10.1016/j.cell.2015.01.029 (2015).
2. Altschul, S. F. et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–402 (1997).
3. Li, Y., He, L., Lucy He, R. et al. A novel fast vector method for genetic sequence comparison. *Sci Rep* **7**, 12226 https://doi.org/10.1038/s41598-017-12493-2 (2017).
4. Deng, M., Yu, C., Liang, Q., He, R. L. & Yau, S. S.-T. A novel method of characterizing genetic sequences: genome space with biological distance and applications*. PLoS ONE***6**, e17293 (2011).
5. Hoang, T., Yin, C. & Yau, S. S.-T. Numerical encoding of dna sequences by chaos game representation with application in similarity comparison. *Genomics* **108**, 134–142 (2016).
6. Ro¨hling S, Linne A, Schellhorn J, Hosseini M, Dencker T, Morgenstern B The number of k-mer matches between two DNA sequences as a function of k and applications to estimate phylogenetic distances. *PLoS ONE* **15(2)**: e0228070. https://doi.org/10.1371/journal.pone.0228070 (2020).
7. Jain, C., Rodriguez-R, L.M., Phillippy, A.M. *et al.*High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9,** 5114 https://doi.org/10.1038/s41467-018-07641-9 (2018).
8. Wong, Karen M. and Suchard, Marc A. and Huelsenbeck, John P. Alignment Uncertainty and Genomic Analysis . *SCIENCE* **319(5862),** 473-476, (2008).
9. Kemena C, Notredame C, Upcoming challenges for multiple sequence alignment methods in the high-throughput era". **25 (19)**: 2455–2465, doi:10.1093/bioinformatics/btp452 (2009).
10. Hoang, T., Yin, C. & Yau, S. S.-T. Numerical encoding of dna sequences by chaos game representation with application in similarity comparison. *Genomics* **108**, 134–142 (2016).
11. Hatje, K. & Kollmar, M. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. *Front Plant Sci***3**, 192 (2012).
12. Leimeister, C.-A. & Morgenstern, B. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics* **30**, 2000–2008 (2014).
13. Zielezinski, A., Girgis, H.Z., Bernard, G. et al. Benchmarking of alignment-free sequence comparison methods. *Genome Biol* **20**, 144 , https://doi.org/10.1186/s13059-019-1755-7, (2019).
14. Jeffrey, H. J. Chaos game representation of gene structure. *Nucleic Acids Research* **18**, 2163–2170 (1990).

15. Deschavanne PJ1, Giron A, Vilain J, Fagot G, Fertil B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol.* **16(10)**.,1391-1399, (1999). DOI:10.1093/oxfordjournals.molbev.a02604
16. Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A. & Fletcher, M. Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics* **17**, 429–437 (2001).
17. Stan, C.P. Cristescu, E.I. Scarlat, Similarity analysis for DNA sequences based on chaos game representation case study: The albumin, *J. Theoret. Biol.,* **267**, 513-518, (2010)
18. Hai ming Ni, Da wei Qi, Hongbo Mu, Applying MSSIM combined chaos game representation to genome sequences analysis, *Genomics* **110 (3)**, 180-190. https://doi.org/10.1016/j.ygeno.2017.09.010 (2018).
19. Lichtblau, D. Alignment-free genomic sequence comparison using FCGR and signal processing. *BMC Bioinformatics* **20**, 742 https://doi.org/10.1186/s12859-019-3330-3, (2019).
20. Jácome, R., Campillo-Balderas, J.A., Ponce de León, S. et al. Sofosbuvir as a potential alternative to treat the SARS-CoV-2 epidemic. *Sci Rep* **10**, 9294 (2020). https://doi.org/10.1038/s41598-020-66440-9

Figures

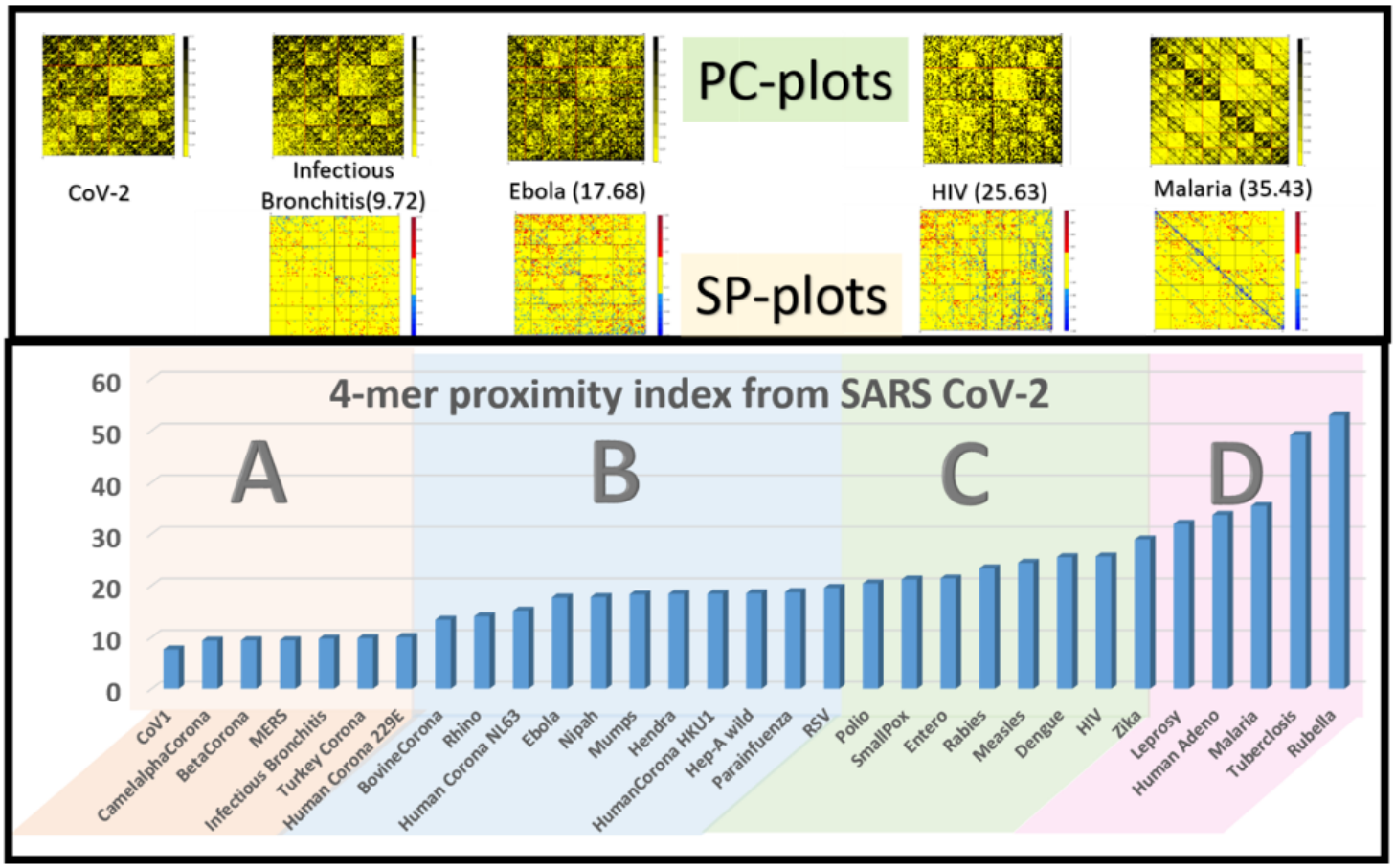

Figure 1

The PC-plot, SP-plot and tetra-nucleotide (4-mer) proximity indices for pathogens paired with SARS-CoV-2. The top panel shows PC-plots (upper half) of five pathogens, SARS-CoV-2 and one each from four categories, and SP-plots (lower half) of four pathogens compared with SARS-CoV-2. The number in parentheses is the tetra-nucleotide proximity index of the corresponding pathogen pair. The bottom panel shows the variation in tetra-nucleotide proximity indices for 31 pathogen pairs. The pathogens are categorized in four categories, A, B, C and D based on their dissimilarity from SARS-CoV-2.

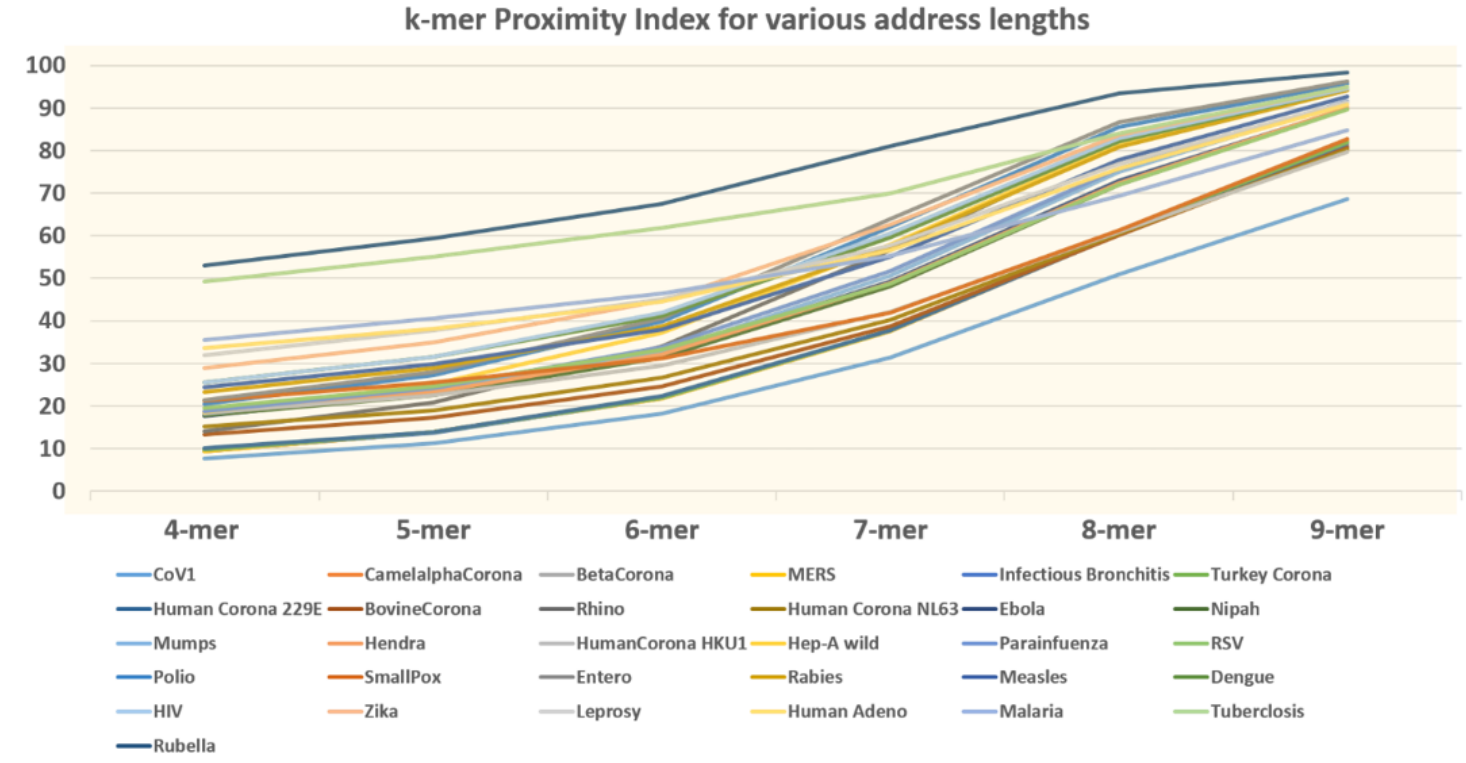

Figure 2

Variation of k-mer proximity index of 31 pathogen pairs with the word length.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile.docx