Chapter

Entropy Based Biological Sequence Study

Bimal Kumar Sarkar

Abstract

SARS-CoV-2 virus strains are taken into consideration for the analysis of digitized sequences of information by means of the notions of entropy. The occurrence of a particular pattern in the corona viral sequence is paid a special attention. The incidence of genetic word is represented in a density means. The incidence frequency of the q-gram genetic word is determined with the help of finite impulse response (FIR) filter along the sequence. It is in turn, used for the determination of the probability distribution of the genetic word incidence as the input for the calculation of entropy in the sequence. The sequence entropy is further used for principal component analysis (PCA) to determine the similarity/dissimilarity between the viral sequences. We have considered seven human corona virus sequences. Entropy based similarity study for SARS-CoV-2 strains is presented in this work.

Keywords: sequences, genetic information, FIR filter, entropy, PCA, corona virus

1. Introduction

The entropy of amino acid sequences in DNA of an organism can be considered as the measure of diversity of proteins. The higher the value of entropy, the greater the possibility of variation in the information content coded by the nucleic acid [1]. This theory is utilized in the present study to understand the variation in the genetic sequences of different novel corona viruses that have infected people across the world leading to one of the world's biggest pandemics. The pandemic itself highlights the importance of tracking the dynamics of viral transmission in real-time. Moreover, as the virus mutates frequently, each sequence is studied and compared with others to understand the variation of information that is transmitted from one species to the other. Hyper-variable genomic hotspot for the novel coronavirus SARS-CoV-2 has already been identified by Wen et al. [2]. Likewise, the similarities in the genetic code would also provide important information in understanding the virus and its prevention.

Corona virus molecule has a single-stranded, positive-sense RNA genome of length of approximately 27 to 32 kilobases (kb). The genome sizes of HCoV-229E and HCoV-NL63 are approximately 27.5 kb, and it is more than 30 kb for HCoV-OC43 and HCoV-HKU1. It possesses the RNA harbors a 50-cap structure and a 30-polyadenylate tail which enable to play a role of messenger RNA (mRNA) [3–10].

This study presents identification and analysis of regions of similarity in SARS-CoV genetic sequence [11–13]. According to information theory, individuality of a

species can be aggregates that propagate information from past to future. The Shannon Entropy is considered as a measure for the order/disorder state of nucleotide sequences of the DNA [14]. The information in a genetic code is comprised of an alphabetic sequence of the four letters A, C, G, and T, which symbolizes the four nucleotides, namely, adenine (A), cytosine (C), guanine (G) and thymine (T). The sequences have been recognized for most of the SARS-CoV-2 genes and are accessible in computer readable form. The probability of occurrence of a combination of a group of symbols in a sequence is the measure of order in a sequence. An alignment free approach of DNA sequence analysis, *n*-mer/word frequency estimation, is attempted in this work.

2. Methodology

Our method is based on the observation through a sliding "counter "of width W over DNA sequence [15]. A certain number of q-grams called as bins are set in the counter. As there are only four letters in the DNA alphabet, viz., {A, C, G, T} the number of all combinations of q-grams in a DNA sequence is 4^q.

Definition 1. q-gram of Sequence.

Given a sequence 'seq', when a window of length q slides over the characters of 'seq', its q-grams are formed. For a sequence 'seq, there are $|seq| - (q - 1)$ *q-grams.*

The number of all possible q-grams or called as "bin" is 4^q. Bins can be arranged in lexicographic order, and b_i is used to denote the i^{th} bin in this order. All the possible bins are denoted as:

$$B_q = \{b_1,\ b_2,\ \ldots\ldots\ b_{4^q}\} \tag{1}$$

Example 1. One-gram bins are $B_1 = \{A,\ C, G, T\}$, consisting 4 bins. Two-gram bins are $B_2 = \{$AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT$\}$, consisting 16 bins.

Definition 2. Bin Signature.

For a sequence, the q-gram bin signature, S_j is a mapping with the bin b_j $\left(b_j \in B_q\right)$ where i^{th} bit in S_j, is corresponding to the presence or absence of b_j. For a sequence 'seq, there are $|seq| - \left(\left|b_j\right| - 1\right)$ bits in S_j.

Example 2. Consider a sequence, S = "AACTCG". Its two-grams (q = 2) signature in the sequence is S_2 = [0 1 0 0 0].

Definition 3. Filter.

A sequence x[n] is filtered through mapping of the sequence into output sequence y[n] via a weighted window b by means of the convolution summation as

$$y[n] = \sum_{i=0}^{k} b_i x[n - i] \tag{2}$$

b is independent of x[n] and y[n], where n is the time index. y[n] is the response of the filter to input signal x[n]. The filter is finite impulse response (FIR) digital filter. The term digital filter arises because it operates on discrete-time signals. Finite impulse response arises because the filter output is computed as a weighted, finite term sum, of past and present (**Figure 1**).

Example 3: Weighted filter output of S_A with the weighted window β = [0.2 0.1 0.3 0.4] is as follows:

S_A = [1 1 0 0 0 0].

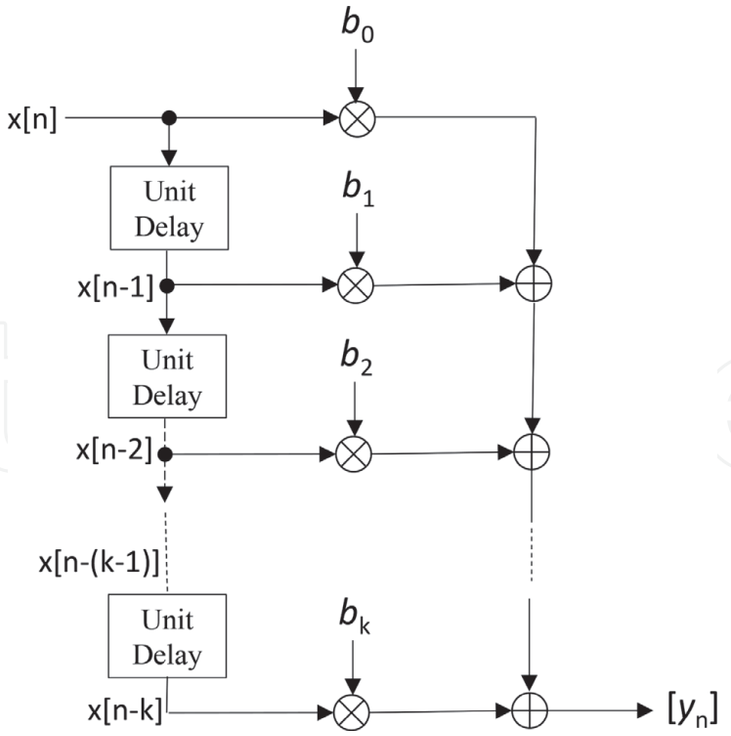

Figure 1.
Block diagram of finite impulse response (FIR) digital filter.

$$y_A[n] = \sum_{0}^{3} \beta_k S_A[n-k] \text{ with } \beta_0 = 0.2, \beta_1 = 0.1, \beta_2 = 0.3, \beta_3 = 0.4.$$

$$\Rightarrow y_A[n] = \beta_0 S_A[n] + \beta_1 S_A[n-1] + \beta_2 S_A[n-2] + \beta_3 S_A[n-3]$$

y_A = [0.2 0.3 0.4 0.7 0.4 0]; Similarly for other nucleotide viz., C, G, T, the output is obtained as,

y_C = [0.2 0.0 0.2 0.1 0.5 0.5]; y_G = [0.0 0.0 0.0 0.0 0.0 0.2]; y_T = [0.0 0.0 0.0 0.2 0.1 0.3];

For nucleotide density calculation, evenly distributed window of unit value is considered. As explained, the output of the convolution summation represents the nucleotide density along the sequence. The detail algorithms for bin construction, bin signature, filter operation is displayed in **Tables 1–3** respectively.

Input: q - length of bin **Output:** set of bins $B_q = \{b_1, b_2, \ldots\ldots b_{4^q}\}$

```
1: 0 ← bincount;
2: 4^q ← n;
3: cell(1,n) ← bin;
4: for first = 1:4 do
5:     for qth = 1:4 do
6:     convert integer to nucleotide character ([first ... qth]) ← binq;
7:     bincount = bincount +1;
8:     binq← bin{bincount};
9:     end
10:  end
11:  bin ← Bq
```

Table 1.
Bin construction.

Input: Sequence (seq), bin (b) **Output:** Bin Signature

```
1: m ← length (seq);
2: nbin← length (b);
3: for i ← 1... .m - (nbin - 1) do
4:    if seq (i: i + nbin-1) = b then
5:       signature (i) = 1
6:    else
7:       signature (i) = 0
8: end
9: signature ← Bin Signature
```

Table 2.
Bin signature.

Input: *BinSignature*, *window* **Output:** *filter*

```
1: w ← length (window);
2: window = 1/w*array of ones(1,w);
3: 0 ← sum
4: for i ← 1... . length (window) do
5: make array of zeros with length of i − 1 ← zero
6: sum = sum + window (i) * array[zeros BinSignature(1:(length(BinSignature)-(i-1)))]
7: end
8: filter ← sum
```

Table 3.
Filter.

3. Sequence analysis

The filter output is taken as a density distribution for DNA sequences. The density distribution is based on q-gram word density, which in turn is considered for the determination of Shannon Entropy as

$$y_i = -\sum_{j=1}^{q} p_{ij} \log p_{ij} \tag{3}$$

where p_{ij} is the probability of appearance of the *j*th genetic letter at *i*th position in the genetic sequence. Further we want to find a similarity/dissimilarity measure between two entropy distributions $\rho_i = (y_{i1}, y_{i2}, \ldots, y_{in})$ and $\rho_j = (y_{j1}, y_{j2}, \ldots, y_{jn})$. We construct the data matrix D comprising elements $[\rho_1, \rho_2, \ldots, \rho_m]'$, where *m* is the number of sequences. Principal Components Analyses (PCA) is used to estimate scores between density distributions such that it reduces multidimensional data sets to lower dimensions with the consistent original data matrix [16].

We determine the dissimilarity between two sequences from the scores in the first three principal components by computing the Euclidean distance between pairs of density distributions in the m-by-n data matrix D. Rows of D correspond to sequence (observations) and columns correspond to position index in the sequence (variables). Thus, Euclidean distance X is a row vector of length *m (m–1)/2*, corresponding to pairs of observations in D. The distances are arranged in the order *(2, 1), (3, 1), ..., (m, 1), (3, 2), ..., (m, 2), ..., (m, m–1)*. X is used as a dissimilarity matrix in clustering or multidimensional scaling. An unweighted pair group method with arithmetic mean (UPGMA) is employed on PC scores for the construction of a phylogenetic tree [17]. UPGMA uses a local objective function to construct a rooted bifurcating tree.

4. Results and discussions

The nucleotide density distribution was obtained through FIR filter. We have calculated the density distribution for one-, two-, three-, gram nucleotide for different species. Secondly we have calculated entropy distributions ρi = (yi1, yi2, … , yin) and ρj = (yj1, yj2, … , yjn). The variation of entropy with position for all other sequences are calculated for the above three combinations. The entropy values were found to be minimum for mono-mer density distributions in individual sequences while increasing linearly for di-mers and codons respectively. Observations based on the position of the n-mers in sequences of SARS-CoV-2 DNA reveals significant minimum entropy regions for codons. **Figure 2** shows the entropy profile calculated over 29000 bases for 7 DNA sequences. Similar analysis profile for mono-mers and di-mers does not show overlapping regions for different sequences. This suggests that codons are more effective in transferring information through different species. Codon bias has been reported for HIV 1 virus [18]. Therefore, it can be inferred that in various novel coronavirus strains, the codons at specific positions are the highest bias representing minimum entropy and hence carry the maximum information. Further studies with the sequences of these loci can be useful genetic engineering for developing vaccines or taking control over the spread of the second wave of the pandemic.

We have chosen seven SARS Corona virus sequences (SARS-COV) from various countries. The details of the organism are presented in **Table 4**.

Based on FIR filtering, firstly the nucleotide density distribution is generated. We have calculated the density distribution for one-, two-, three-, gram nucleotide for different species. Secondly we have calculated entropy distributions ρi = (yi1, yi2, … , yin) and ρj = (yj1, yj2, … , yjn). **Figure 3** displays the spatial variation of the entropy along the SARS-COV sequence for seven species.

In fact it is inconvenient to realize all the entropy variation in 2D graphical representation. For example, the organism HKU1 shows the positions where it possesses the minima in entropy values. Some are demonstrated at the positions, around 7400, 10000, 23000 etc. the Amsterdam strain, NL63 has shown minima at around 7300, 8000 etc. But other strains exhibit their entropy representation in a crowded manner. It is difficult to understand the variation for them differentially. Rather it is more comprehensive to show the entropy variation for all sequences (total 7) in a single panel. It has been shown in **Figure 3**.

The present work intends to assess the variability and complexity at each nucleotide site with the calculation of entropy for each position using the Shannon entropy formula, Eq. (2). The low entropy regions around 7400 and 9000 position

Sequence No.	Strain	Accession No.	Place
1	Wuhan-Hu-1	MN908947	Wuhan
2	CV7	DQ898174	Canada
3	MERS-CoV /C1272	MH734115	Kenya
4	HCoV-OC43	KU131570	UK/London
5	NL63	DQ445912	Amsterdam
6	HCoV_229E	MN306046	Seattle/USA
7	HKU1	MH940245	Thailand

Table 4.
SARS-COV strains with their complete genome sequence, accession no. and source.

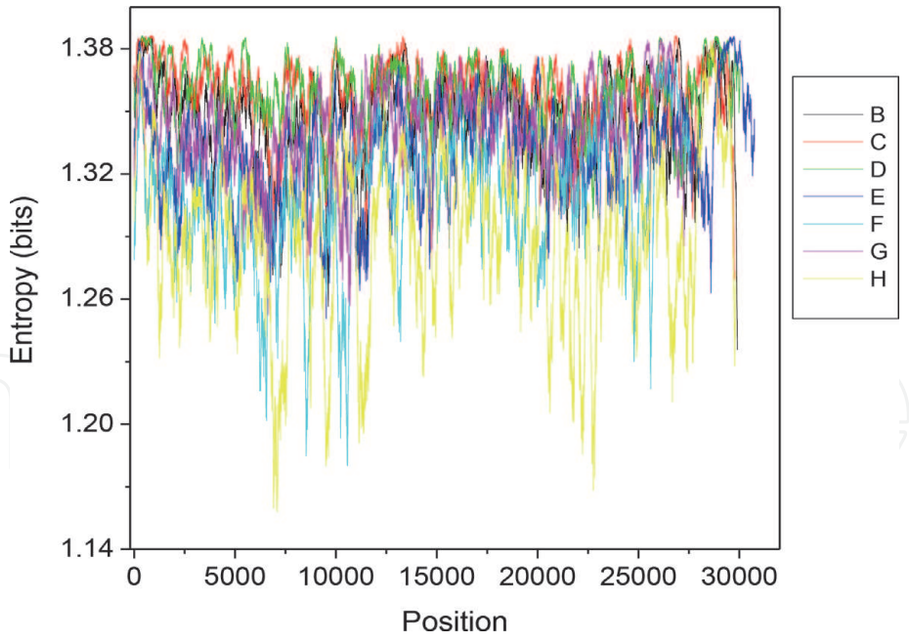

Figure 2.
Entropy profile of seven SARS-COV sequence. Entropy is calculated based on single nucleotide distribution. Sequences are: B: Wuhan-Hu-1; C CV7; D: MERS-CoV/C1272; E: HCoV-OC43; F: NL63; G: HCoV_229E; H: HKU1 (see ***Table 4****).*

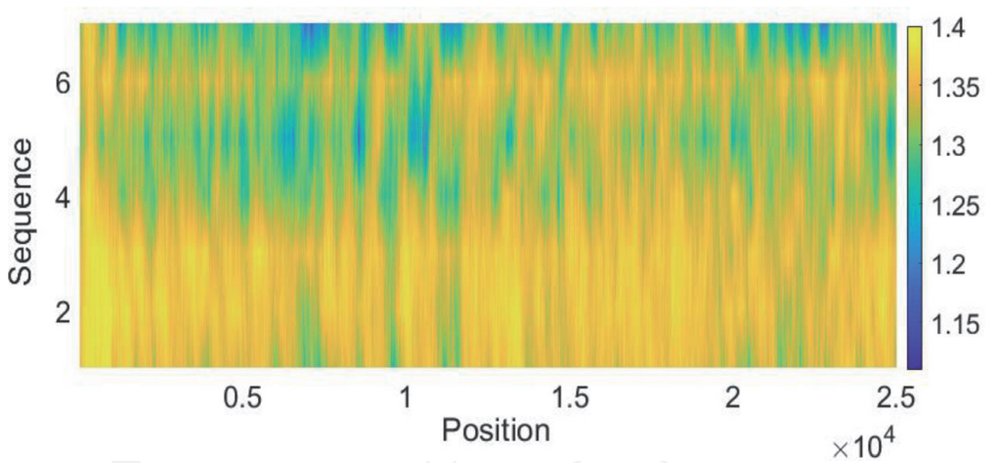

Figure 3.
Entropy profile of 7 SARS-COV sequences. Entropy is calculated based on single nucleotide distribution. Sequences are represented as number starting from 1 to 7. (see ***Table 4****).*

are common to all 7 sequences (**Figure 3**). Entropy (Y_i) is an important parameter for the understanding of sequential stability. Y*i* becomes maximal when all symbols occur at equal probability. On the other hand, Y*i* becomes the least if one symbol occurs at probability 1 and in that case the other symbols will be forbidden. It means that lower the value of entropy the site is more stable without much complexity. Under this assumption, the zone around the site 7400 and 9000 position are most stable for all strain/species. It may find a good structural relationship between the regions of low entropy and the secondary structure of proteins which include α-helix, β-sheets and loops regions.

Strain no. 4–7 (HCoV-OC43; F: NL63; G: HCoV_229E; H: HKU1) show the stability with lower entropy around 8 K, 9 K, 11 K, 12 K site position. But this behavoiu is not exhibited in case of the strains numbers 1–3 (Wuhan-Hu-1; C CV7; D: MERS-CoV/C1272). If one can go through these strains, as a whole, it is noticed that the entropy is increasing or in turn the complexity is more. It is an indication of

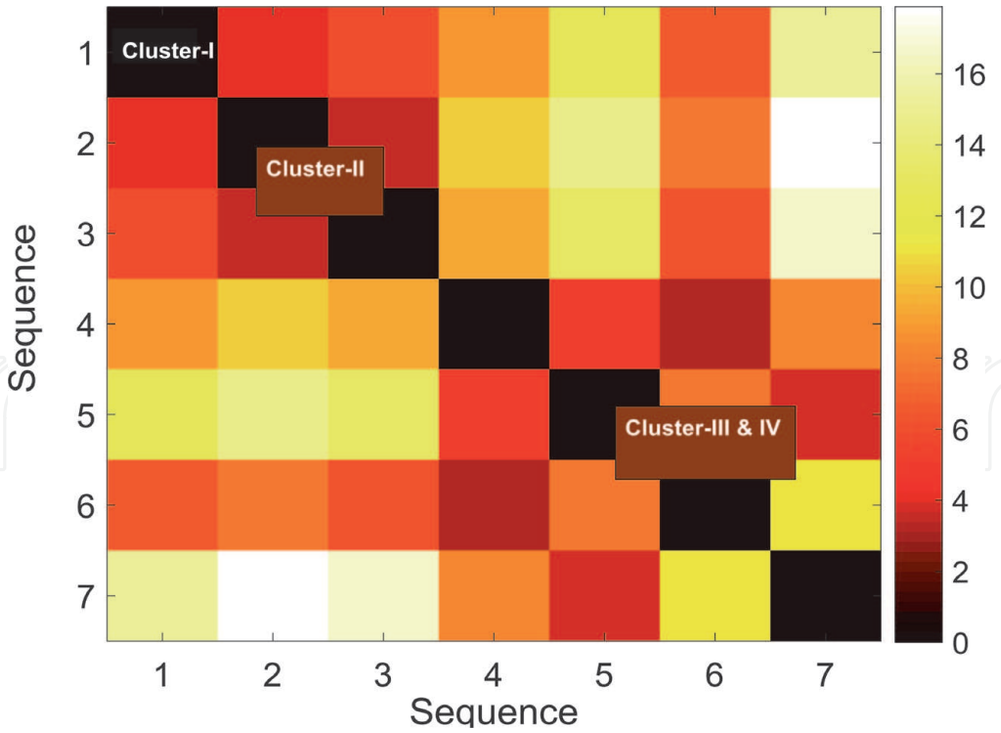

Figure 4.
Dissimilarity matrix for 7 SARS-COV sequences.

evolutionary development among the SARS-COV strains. Based on site entropy we prepared the dissimilarity matrix for the sequences (**Figure 4**).

The dissimilarity matrix demonstrates the existence of 4 different clusters. One can see that the SARS-COV sequences in a cluster shows less dissimilarity among themselves. In other way to mean that the sequences have much similarity residing in a cluster [19]. The COVID sequence appearing in cluster I is typically from Wuhan, China. The Wuhan virus genome sequence examination found β-CoV strain [20]. The Wuhan novel β-CoV revealed 88% similarity with the sequence of two bat-derived SARS-COV, bat-SL-CoVZC45 and it was named "SARS-CoV-2" by the International Virus Classification Commission. The genome of SARS-CoV-2 sequence has the similarity with the typical CoVs. It encompasses more than ten open reading frames (ORFs). The first ORFs covers about two-thirds of viral RNA, which get translated into two large polyproteins, pp1a and pp1ab. These proteins assist to form the viral replicase transcriptase complex [21]. The remaining one-third of viral RNA take part in translation of four structural proteins: spike (S), envelope (E), nucleocapsid (N) and membrane (M) proteins [22].

Cluster-II comprising of two strains CV7, MERS-CoV, belong to β-CoV genera, which also includes SARS-CoV-2 strain as placed singly in cluster-I. Two HCoVs of strains HCoV-229E and HCoV-OC43 being placed in the mixed Cluster of III and IV, are the members of α-CoV genera. From the cluster presentation (**Figure 5**), it will be understood that they belong to cluster-III. Remaining two strains, NL63 and HKU1 are placed in cluster IV.

Phylogenic relation among the strains is represented in **Figure 6**. We obtain the phylogenetic tree of the data set based on unweighted pair group method with arithmetic mean (UPGMA) on PC scores. Phylogenetic tree analysis clearly shows the relationship among all COVID strains under each cluster. We further sub-cluster in each cluster based on their genetic distance (GD). We have considered PC score to determine the dissimilarity or genetic distance between two organisms.

Explicitly the COVID strains are placed in a cluster description (**Figure 5**). The scores are determined in the principal component analysis. Three principal

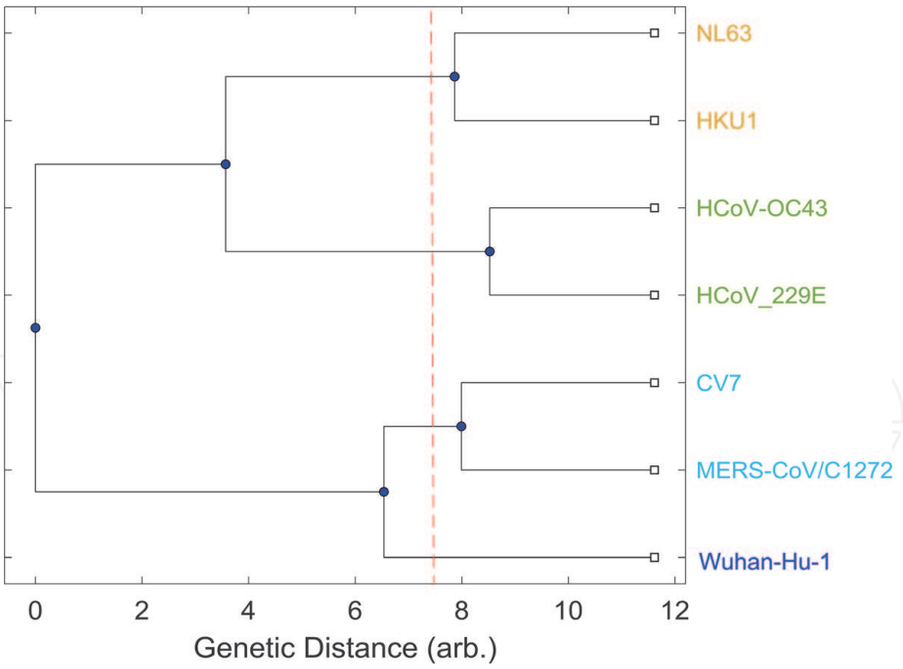

Figure 5.
Scatter plot PC values for 7 SARS-COV sequences: Cluster-I (Wuhan-Hu-1) is encircled with deep blue color). Cluster-II (CV7 and MERS-CoV) is encircled with light blue color. Cluster-III (HCoV-229E and HCoV-OC43) is encircled with green color. Cluster-IV (NL63 and HKU1) is encircled with yellow color.

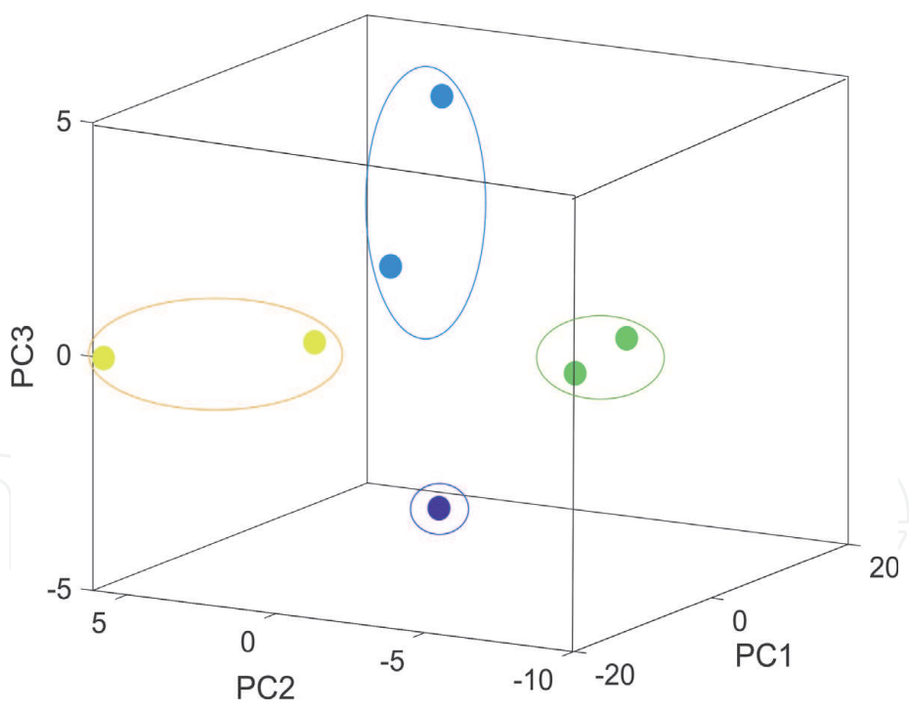

Figure 6.
The phylogenetic tree of 7 SARS-COV sequences.

components are taken into consideration. Each strain is represented as state point by scatter plot in the three PC space. Cluster presentation is well agreement with phylogenic relations. Wuhan-Hu-1 strain is well isolated from all other strains. It belongs to cluster-I. Each of other three clusters possess two-member strain. Cluster-II comprises of two strains CV7 and MERS-CoV belonging to β-CoV genera (encircled with blue color ellipse in **Figure 5**). Already it is mentioned in the previous section that the strains HCoV-229E and HCoV-OC43 exist in Cluster of III.

It is displayed by two state points encircled in green colored ellipse. Remaining pair of strains, NL63 and HKU1 are placed in cluster IV which is marked by yellow colored ellipse.

5. Conclusions

The entropy has been used to select SARS-COV genome regions for stability zone detection. Even though a great deal of genetic variation is generally found, the present entropy calculation is sufficient to observe low informational complexity regions, which are representation of the conserved sites of the sequence. The low entropy regions are related to important functional domains in the proteins of these viruses. Based on entropy calculations seven SARS-COV genomes have been phylogenically described. The clusters of the genome formation is well understood.

Acknowledgements

The authors are thankful to the Department of Physics, Adamas University for providing computational facility. The authors acknowledge the collection of sequence data from NCBI gene bank.

Author details

Bimal Kumar Sarkar
Department of Physics, Adamas University, Kolkata, India

*Address all correspondence to: bks@physics.org.in

References

[1] Hasegawa, M. and Yano, T.A., The genetic code and the entropy of protein. Mathematical Biosciences, 24(1–2), 169–182 (1975).

[2] Wen, F., Yu, H., Guo, J., Li, Y., Luo, K. and Huang, S., Identification of the hyper-variable genomic hotspot for the novel coronavirus SARS-CoV-2. The Journal of Infection, 80 (60), 671–693 (2020).

[3] P. Zhou, X.L. Yang, X.G. Wang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 579, 270–273 (2020).

[4] E. de Wit, N. van Doremalen, D. Falzarano, et al., SARS and MERS: recent insights into emerging coronaviruses. Nat. Rev. Microbiol., 14, 523e534 (2016).

[5] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet, 395 (10225), 689–697 (2020).

[6] S. Su, G. Wong, W. Shi, et al., Epidemiology, Genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol., 24, 490e502 (2016).

[7] S. Perlman, J. Netland Coronaviruses post-SARS: update on replication and pathogenesis, Nat. Rev. Microbiol., 7, 439e450 (2009).

[8] R. Lu, X. Zhao, J. Li, et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet, 395, (10224) 565–574 (2020).

[9] A.R. Fehr, S. Perlman, Coronaviruses: an overview of their replication and pathogenesis. Methods Mol. Biol,, 1282, 1e23 (2015).

[10] P.S. Masters, The molecular biology of coronaviruses. Adv. Virus Res, 66, 193e292 (2006).

[11] X.Y. Ge, J.L. Li, X.L. Yang, et al., Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature, 503, 535e538 (2013).

[12] Chinese SARS Molecular Epidemiology Consortium, Chinese Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science, 303, 1666e1669 (2004).

[13] V.S. Raj, H. Mou, S.L. Smits, et al., Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. Nature, 495, 251e254 (2013).

[14] Schmitt, A.O. and Herzel, H., Estimating the entropy of DNA sequences. Journal of theoretical biology, 188(3), 369–377 (1997).

[15] Saha, P. and Sarkar, B.K., Entropy based analysis of genetic information. Journal of Physics: Conference Series, 1579 (1), 012003 (2020).

[16] Novembre, J., Stephens, M., Interpreting principal component analyses of spatial population genetic variation. Nat Genet., 40, 646–649 (2008).

[17] JF Yu, JH Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation. Commun Math Comput Chem, 63, 493–512 (2010).

[18] Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R., Codon catalog usage is a genome strategy modulated for gene expressivity. Nucl Acids Res., 9 (1), 213–213 (1981).

[19] Chan J F-W, Yuan S., Kok K-H., Wang K-KChu H., A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet, 395 (10223), 514–523 (2020).

[20] Lu R., Zhao X., Li J., Niu P., Yang B., Wu H., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet, 395 (10224), 565–574 (2020.

[21] A.R. Fehr, S. Perlman, Coronaviruses: an overview of their replication and pathogenesis. Methods Mol. Biol., 1282, 1e23 (2015).

[22] K. Knoops, M. Kikkert, S.H. Worm, et al., SARS-coronavirus replication is supported by a reticulovesicular network of modified endoplasmic reticulum. PLoS Biol., 6, e226 (2008).