Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Genomics 113 (2021) 778–784

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Clustering and classification of virus sequence through music communication protocol and wavelet transform

Tirthankar Paul [a,*], Seppo Vainio [b], Juha Roning [a]

[a] InfoTech Oulu, Biomimetics and Intelligent Systems Group (BISG), Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland
[b] InfoTech Oulu, Faculty of Biochemistry and Molecular Medicine, Biocenter Oulu, Laboratory of Development Biology, University of Oulu, Oulu, Finland

ARTICLE INFO

Keywords:
Coronavirus
Haar wavelet
SVM
Protein music
MIDI

ABSTRACT

The coronavirus pandemic became a major risk in global public health. The outbreak is caused by SARS-CoV-2, a member of the coronavirus family. Though the images of the virus are familiar to us, in the present study, an attempt is made to hear the coronavirus by translating its protein spike into audio sequences. The musical features such as pitch, timbre, volume and duration are mapped based on the coronavirus protein sequence. Three different viruses Influenza, Ebola and Coronavirus were studied and compared through their auditory virus sequences by implementing Haar wavelet transform. The sonification of the coronavirus benefits in understanding the protein structures by enhancing the hidden features. Further, it makes a clear difference in the representation of coronavirus compared with other viruses, which will help in various research works related to virus sequence. This evolves as a simplified and novel way of representing the conventional computational methods.

1. Introduction

The year 2020 begins with a threat of coronavirus originated in China, and later spread to the rest of the world. The novel coronavirus has reached almost every country in the world. Multiple numbers of pneumonia cases were noticed in Wuhan city, Hubei Province, China, in December 2019. Later, the disease was recognised as a novel coronavirus. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) is the seventh family member of the coronavirus family [1,2]. Initially, the cases were at first reported from China, later in Japan, South Korea, Singapore Thailand, mostly Asian countries. Gradually, the cases were found in Europe and America [3]. The virus is affecting people around the globe, and almost all countries are affected by the virus. More than seven million people have been affected, and around four hundred fifty thousand people died due to this virus by 13th June 2020. In this paper, the coronavirus protein sequence was translated into music through MIDI protocol and compared with intra-family and inter-family members. Representing genome data in a non-conventional way has always been appreciated by the researches; for example, the genome sequence was portrayed as an image, based on chaos game representation to analysis different biological features in the recent publications [4–6]. Here, A set of 56 virus protein sequences among Coronavirus, Influenza and Ebola were studied and classified through their auditory pattern.

The SARS-CoV-2, which causes a severe respiratory syndrome, is a positive RNA strand [7]. World Health Organization (WHO) enlisted unknown caused an epidemic and pandemic potential disease such as the Middle East respiratory syndrome (MERS), Disease X and severe acute respiratory syndrome (SARS) in their priority list of pathogens in April 2018 [8]. Initially, the disease was suspected as a Disease X aetiology by WHO [9]. But soon after it was denoted a novel coronavirus (2019-nCoV) caused disease COVID-19 by WHO [10,11]. In the study, it was found that the 2019-nCov has 79.5% protein sequence similarity with SARS-CoV and 96% similarity with SL-Cov-RaTG13, known as bat coronavirus [12]. The Chinese group of virologists renamed the virus HCoV-19 [13], while the internationally, Coronavirus Study Group (CSG) renamed it SARS-CoV-2 [14].

It is not a new trend to represent a species, DNA, RNA or genome into musical form. Japanese scientist Susumo Ohno showed a connection between reoccurrence of repeat unit and musical repetitions [15]. Initially, the work Biomusic was translated into musical notes from four DNA bases which gave an auditory pattern of DNA sequence but lacked a rhythmic and musical point of view [16]. Later, many approaches were attempted like codon reading and physical properties of DNA bases, to translate DNA nucleotides into music [17,18]. There are few studies

* Corresponding author.
E-mail addresses: tirthankar.paul@oulu.fi (T. Paul), seppo.vainio@oulu.fi (S. Vainio), juha.roning@oulu.fi (J. Roning).

https://doi.org/10.1016/j.ygeno.2020.10.009
Received 2 August 2020; Accepted 13 October 2020
Available online 16 October 2020

Table 1
Genome sequence clustering algorithms.

Reference	Algorithm	Significance	Year
[32]	Natural Vector (NV)	DNA sequences into twelve statistical points vectors.	2011
[30]	mBKM with DMk	The occurrence and position of k-tuples of DNA sequences	2012
[33]	Linclust	Reduce the time complexity for the large dataset	2018
[34]	MeShClust	Clustering method based on shift algorithm of image processing.	2018
[35]	Haar wavelet filtering	Detecting cancerous and non-cancerous genome.	2018
[31]	Accumulated Natural Vector (ANV)	DNA sequences into eighteen statistical points vectors.	2019

based on protein music where the protein was translated into music. The study showed that the secondary structure of proteins could be converted into a musical sequence [19]. DNA music was also formed by mapping amino acid into different pitches. Therefore, 20 amino acid are assigned to 20 different notes and the protein music can be created [20]. This kind of musical transformation would identify the difference between protein folding and amino acids in terms of understanding the facts to regulate the cell process [21]. DNA music can also be created based on the presence of short tandem repeats (STR) in a CODIS region, making it very unique for the individuals [22]. The STR sequence and the STR frequency data were converted as a musical element and performed in musical instrument and digital interference (MIDI) format to make the music melodious. The MIDI is a commonly used protocol to make communication between musical device and computer, mostly in the music industries. Initially, it was developed to create polyphonic music [23]. A musical translation from protein fold can also be a useful method for comparative study of different genome sequences [24]. Musical mapping is not only limited to biological data, but mobile keystroke data could also be mapped into music to secure the credentials [25]. Recently, a musical work from COVID 19 was created by translating protein into music in Massachusetts Institute of Technology, USA [26]. Musical notes can be displayed in several way. Most common way to present the music in a musical scale is with octave scale, diatonic scale, Tone scale and chromatic scale [15,20,21,27,28]. Our study shows auditory representation of virus. Sometime, the audio sequence sounds like music, but no musical scale was followed for the sonification. Therefore, the audio sequence of the virus was presented in form of piano sound [22,29].

Biological studies demand high-performance data analysis. Manual data analysis of genome or long nucleotide sequences may not be efficient for classification or clustering, so machine algorithms are always preferred in this case. Clustering is one method among the others in the process, which plays a very important role in bioinformatics studies. A Most common way to show the clustering among the genomic sequence is phylogenetic analysis. An algorithm based on occurrence and position of k-tuples of DNA sequences was introduced for phylogenetic clustering [30]. A similar type method (Accumulated Natural Vector (ANV)) transforms the DNA sequence into eighteen data points, including nucleotides covariance and distribution [31]. This ANV is the advanced version of the Natural Vector (NV) method which translates the DNA into twelve points and claims to be the most accurate method of clustering [31,32]. Genomic datasets are large scale structure of proteins/nucleotides. The complexity increases when a large set of data 'N' items are to be clustered to a large number 'K' cluster. 'Linclust' algorithms reduce the time complexity for the large dataset clustering [33]. Another algorithm 'MeShClust' was introduced for the classification of DNA sequence-based on mean shift algorithm of image processing study [34]. Sequence comparing the used discrete wavelet transform (DWT) was performed by extracting the k-mers from the genome sequences. The k-mers were mapped and transformed into discrete wavelet to get a

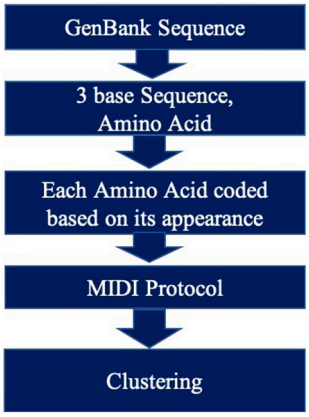


Fig. 1. Steps of the musical conversion.

numeric featured vector for the clustering [35]. A Haar wavelet filtering method was used to decompose the sequences for detecting cancerous genome by Liu et al. [36]. The author extracted statistical data of cancerous and non-cancerous genome and classified via machine learning [36]. Also, Haar wavelet is capable of identifying the short tandem repeats (STR) in a DNA sequence [37]. There is also a standard useful tool, k-means algorithm which is an easy to apply clustering algorithm for genome sequences [38,39]. The above clustering algorithms and their significances are summarized in Table 1. Also, the data can be classified using the support vector machine (SVM). Nucleotide sequences of different species were classified with the SVM model and reconstructed partitioning in Euclidean hyperplanes [40]. An SVM model was applied in a virus genome sequence and achieved a low mean error rate to classify the sequences [41]. Numerical representation of DNA-binding protein sequences was applied for the predicting and classifying the sequences in SVM classifier based on protein properties and features transformation methods [42].

In this study, we created auditory representation of coronaviruses, Influenza and Ebola virus. Our assumption is that the sound pattern could help the researchers to find the virus protein by searching an acoustic sequence. It is natural for a human being to be attracted towards music. In our society, many people have a good knowledge about music and the human brain has an excellent analysing power of sound. Our minds can identify the sound features such as pitch, timbre, volume, rhythm and melody. Translating genome sequence into audio sequence could opens a new door of hidden features of genome data in the form of sound sequence.

2. Methodology

The outlines of the work towards music is shown in Fig. 1. The RNA sequences of SARS-CoV-2 are found in the NCBI database [43]. The genome sequence of the virus has been assigned in GenBank with an accession number MN908947 [10]. The other family members of the virus (SARS-CoV) sequence are assigned in GenBank with an accession number AY278741, having 29,727 nucleotides base [44]. Similarly, the Middle East respiratory syndrome (MERS) coronavirus can also be found in GenBank, and its accession number is KT006149 [45]. In this paper, the musical conversion was performed in the Matlab environment. The programming script was written in Matlab to map the RNA sequence into music. A MIDI toolbox was installed in the Matlab platform to translate nucleotide data to musical elements [29]. The Algorithm 1 is designed to a) download RNA sequence, b) count the protein present in the sequence and c) MIDI musical mapping. The methodology is adopted from the author previous publication [22].

Algorithm 1: Classifying virus families from its audio sequence

access: Virus protein sequence in GenBank.
input: Accession Number in GenBank.
for K value for 3
K-mer count with each possible Amino acid.
Save the three-mer information.
end
for each K-mer (K=3)
Repetition finds
Count Amino acid presence
find highest count for responsible K-mer.
end.
Protein information assigned for Music features.
Set Maximum and Minimum pitch value.
Set Maximum and Minimum duration of note.
Normalised pitch value of each K-mer.
Set MIDI pitch.
Normalised duration K-mer.
Create MIDI note matrix.
Set(:,1) first column 1.
Set(:,2) second column 1.
Set(:,3) third column pitch.
Set(:,4) fourth column volume.
Set(:,5) fifth column note on.
Set(:,6) sixth column note off.
Save the matrix
Matrix to MIDI
Clustering and Classification Pitch value sequences

First, the nucleotide sequence was downloaded from the open-source database NCBI. Then the sequence was converted into a numerical and protein sequences for further process in Matlab. The K-mer analysis was examined and the amino acids were mapped into music. Therefore, the size of the k-mer would be three base codons. So, the codon or amino acid may appear one or many times in the sequence. Each virus sequence has a distinct set of a protein. The number of proteins in the sequence differs with the type of virus. The protein was coded according to the number of occurrences in the sequence. A protein is replaced by a number which indicates the total number of the particular protein present in the sequence. The sequence length and each protein presence take a vital role in numerical mapping of the protein sequence. Later, the musical transformation will make a noticeable difference in the magnitude of the sequence. For example, MADADDAAA is a protein sequence. Total number of 'M' = 1, total number of 'A' = 5, total number of 'D' = 3. So, the coded sequence will be 153533555.

Generally, music has seven different elements, i.e. pitch, volume, timbre, duration, form and texture. Here, pitch and duration were coded based on the physical appearance of amino acids. Also, volume, form and timbre were modulated according to the sequence. The communication message will be assigned a data format/ MIDI matrix to musical devices from the computer [23]. This matrix is the size of N*6 elements, where 'N' is the number of notes. In the MIDI file (M_{ps}), the 'N' row represents a note event, and the 6 columns define different features such as track number, MIDI channel, note value or MIDI pitch, volume, note starting time and note ending time of the MIDI events. Here, track (tn_{ps}) and channel number (cn_{ps}) for piano were set to 1. A constant volume (v_{ps}) for all the note events (N) was fixed to 75. The third column is for MIDI pitch ('mp_{ps}'). The MIDI pitch is to define from the coded frequency ('f') which was mapped from the nucleotide sequence. A small difference in frequency ('f') will make a prominent difference in MIDI pitch (mp_{ps}), for the scaling factor log_2. The MIDI pitch conversion is shown in Eq. (1).

$$mp_{ps} = 69 + 12log_2\left(\frac{f}{440}\right) \tag{1}$$

The fifth column is 'start time' (st_{ps}), and the sixth one is the 'end time' (et_{ps}) of the note and both columns combinedly represent the duration (D_{ps}) of the note. Music can be created by using the set of data as a form of note matrix in the Matlab platform with Ken Schutte MIDI toolbox [29]. The MIDI matrix can be defined as Eq. (2).

$$M_{ps} = \left[\vec{tn}_{ps}, \vec{cn}_{ps}, \vec{mp}_{ps}, \vec{v}_{ps}, \vec{st}_{ps}, \vec{et}_{ps}\right] \tag{2}$$

As described earlier $\vec{tn}_{ps} = \vec{cn}_{ps} = 1$, volume ($\vec{v}_{ps}$) =75, $\vec{mp}_{ps}$ and D_{ps} are mapped based on the protein sequences. Where $D_{ps} = \left|\vec{et}_{ps} - \vec{st}_{ps}\right|$. The M_{ps} was processed through matrix2midi module to generate the audio file of the protein sequences [29].

The pitch values from the MIDI file were taken for finding the similarities among the virus sequences. In this purpose, Euclidian distances were measured to compare the audio signals. The lengths of the signals should be the same to find the distance. However, it is not common to fetch same-length sequence of different species for the study. Here, in this study, three groups of virus, Influenza, Ebola and Coronavirus were examined. There are total of 56 viruses (Influenza, Ebola and Coronavirus) genome sequences were taken for the purpose, and they were downloaded from NCBI database. These three groups of virus have different lengths and sequences, and their auditory translations were filtered through discrete wavelet transform (DWT). A DWT or more precisely Haar wavelet, is fast in computing with reversible lossless transform, and most importantly memory efficient to compare and detect the genome sequence, as shown in the previous studies [36,37]. Coefficients related with Haar wavelet provide the low and high frequency as well as location information in a form of approximate coefficients and detail coefficients of the signal/sequences, respectively [37]. The approximate coefficients and detail coefficients can be explained by the Eq. (3) and (4). The mean value and standard deviation (SD) were obtained from the wavelet transformation to classify the viruses into different clusters. These statistical data with the accession number of the 56 viruses are given in the dataset section.

$$filter(n)_{low} = \sum_{k=-\infty}^{\infty} s[k]g[2.n-k] \tag{3}$$

$$filter(n)_{high} = \sum_{k=-\infty}^{\infty} s[k]h[2.n-k] \tag{4}$$

where, $filter(n)_{low}$ and $filter(n)_{high}$ are the output of low-pass and high-pass filter respectively. s[k] is the protein sequence in numerical form. The low-pass filter coefficient is g(n) and h(n) is the high-pass filter coefficient.

Bayesian optimization model was used to get a good low loss in cross-validation for coronavirus and non-coronavirus data in Fig. 2 [46].

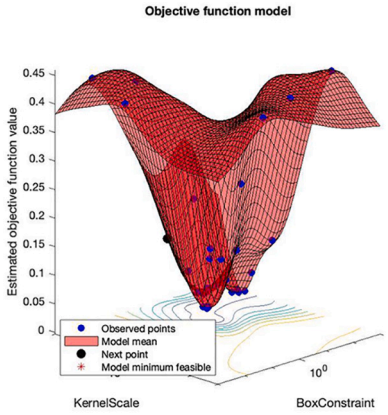


Fig. 2. Cross-validation optimization fit for coronavirus and non-coronavirus data.

Table 2
Accession numbe of whole genome virus sequences.

Reference	Accession number	Description
[47]	AF304460, AY994055, DQ811787, GQ477367, EU420138, EU420139, AF353511, DQ648858, AY585228, DQ011855, AY597011, FJ938068, AY278741, DQ412042, AY304486, M95169, EU022526, EU111742	Coronavirus Family
[48]	KU922529, KU922531, KU922536, KU922542, KY888158	H1N1 Virus from Kerala, India
[49]	NC_014373, NC_004161, NC_006432, NC_014372, NC_002549	Ebolavirus
[43]	NC_003045, NC_004718, KX722529, AB257344, AY291451, AY310120, AY338174, DQ182595, HQ890541	Coronavirus
[43]	AF455734, AF455726, MF955665, AF250130, AF250129, AB731584, CY043013, CY042542, CY020868, NC_002023	Influenza
[43]	KU182909, KY007523, KY007522, KT725391, KT725389, KT725378	Ebolavirus

3. Dataset

The GenBank with an accession number of the genomes, which were not mentioned on the previous paragraph but studied in this paper, are given in Table 2.

4. Result

The acoustic output of a genome sequence may reveal and open many new hidden features which cannot be seen in detail in microscopic images. The musical elements of coronavirus protein music may create a potential impact on our mind. Here, the music conversions were played with a piano instrumental sound. The piano sound of the RNA sequence of SARS-CoV-2 was plotted and shown in Fig. 3a. Similarly, Fig. 3b and Fig. 3c represents the music for SARS-CoV and MERS-CoV, respectively.

The Euclidian distances were measured to show the similarities and dissimilarities among three sequences (SARS-CoV, SARS-CoV-2 and MERS-CoV). The protein sequence and music sequence distance are shown in the upper triangular matrix (marked in yellow) and lower triangular matrix (marked in green) respectively in Table 3. The distance matrix shows distance of 224.0603 for the nucleotide sequence. For more than twenty-nine thousand bases, the Euclidian distance between the two virus sequences is negligible. On the other hand, the Euclidian distance increased to 1534 for the sequence when it was transformed into music. The same trend in Euclidian distance was found for MERS-CoV with SARS-CoV and SARS-CoV-2. The distance is much higher in the music sequence of coronavirus rather than the protein sequence. As a result, the converted music sequence can show a noticeable difference from two very similar nucleotide sequences. This distance can be measured for intra-family members, or those sequence lengths are almost the same. The two sequences need to be the same length to measure Euclidian distance. Influenza, Ebola and Coronavirus are come from different families and also have a vast range of sequence lengths. A Haar wavelet transformation was applied to obtain the statistical values sequences. The viruses were clustered based on the statistical values through K-means clustering algorithm. Fig. 4 and Fig. 5 are the clustered output of virus genome sequences and virus audio sequences. In Fig. 4, most of the virus sequences are placed close to each other, and the clustering algorithm was not significant to show the variation among different group of viruses. On the other hand, the K-means algorithm was applied into the virus pitch value sequences that show three distinct clusters of Influenza, Ebola and Coronavirus in Fig. 5.

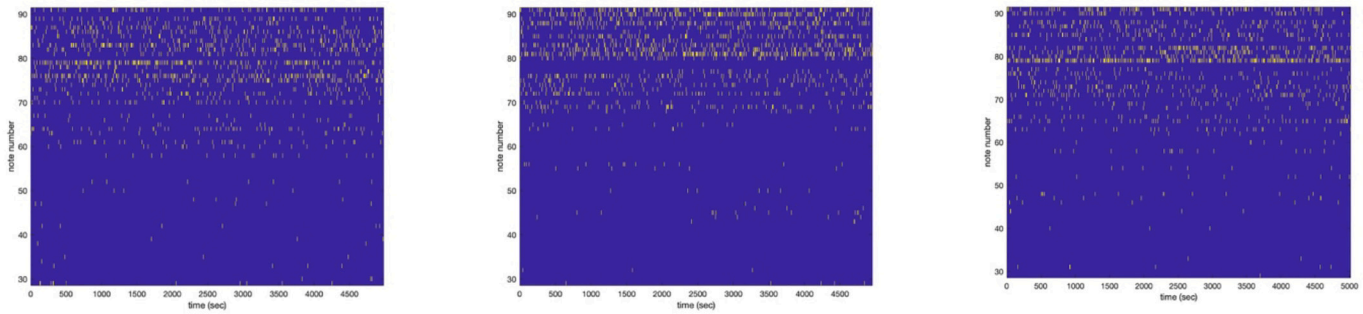
The average detail coefficient from Haar wavelet, of the viruses, are plotted in the boxplot in Fig. 6 (a and b). The boxes lay almost on the same height for genome sequence in Fig. 6a. Differently, the difference of box heights can be shown in Fig. 6b, for audio sequences of the viruses. The audio sequence was created based on the physical features of the protein sequence. Therefore, a small difference in protein sequences creates a considerable change in the statistical values of music sequences. Also, the difference can be visualized in the classifier in Fig. 7, where coronavirus and non-coronavirus audio data were classified with zero loss. On the other hand, the loss of optimized genome sequences data was recorded 0.0377 for the viruses. The classification result improves from genome (Fig. 7a) to audio sequences (Fig. 7b) of coronavirus and non-coronavirus. Therefore, the audio translation of the virus protein sequences enhances the hidden features, which can be identified in the form of a sound signal.

5. Conclusions

This work suggests a way where the Influenza, Ebola and Coronavirus protein sequences can be a sound sequence instead of visual data. The auditory representation of the coronaviruses can help researchers to understand the protein structures in a different way. Sometimes, the

Table 3
Euclidian distance of coronavirus before and after translating into music.

Coronavirus	SARS-CoV	SARS-CoV-2	MERS-CoV
SARS-CoV	X	224.0603	276.4381
SARS-CoV-2	1534	X	290.0707
MERS-CoV	1588.9	1650.4	X



a. SARS-CoV-2 b. SARS-CoV c. MERS-CoV.

Fig. 3. Piano roll plot of virus protein sequence.

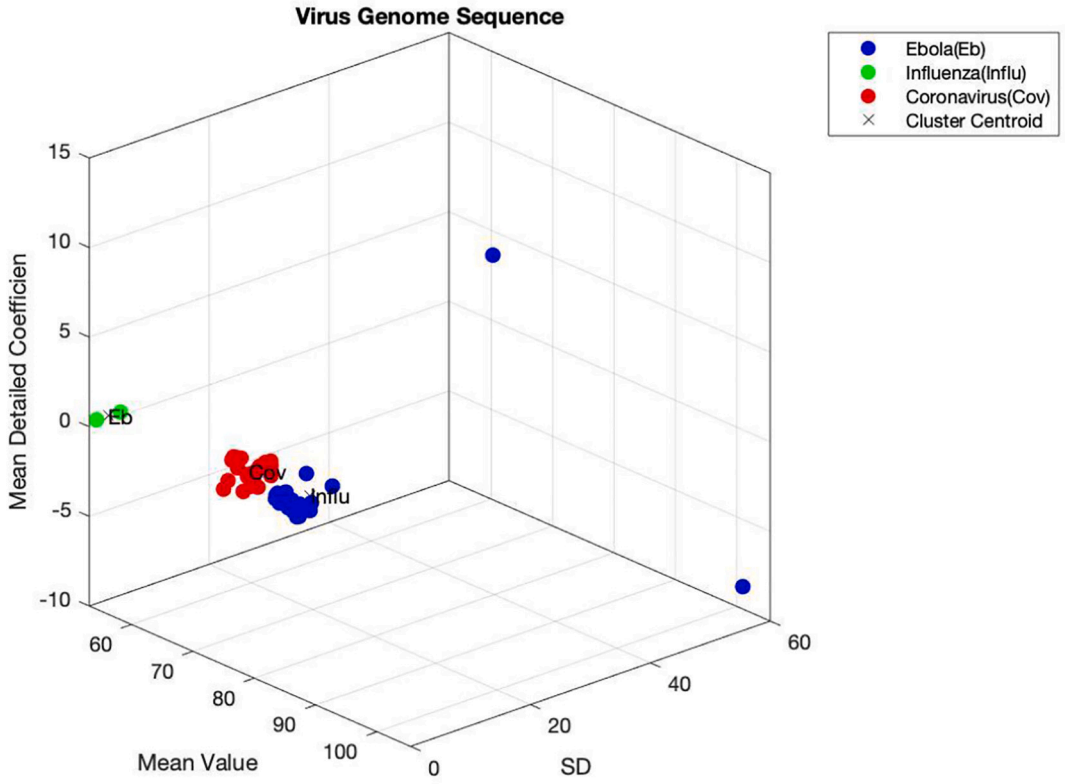


Fig. 4. Clustering of virus sequences.

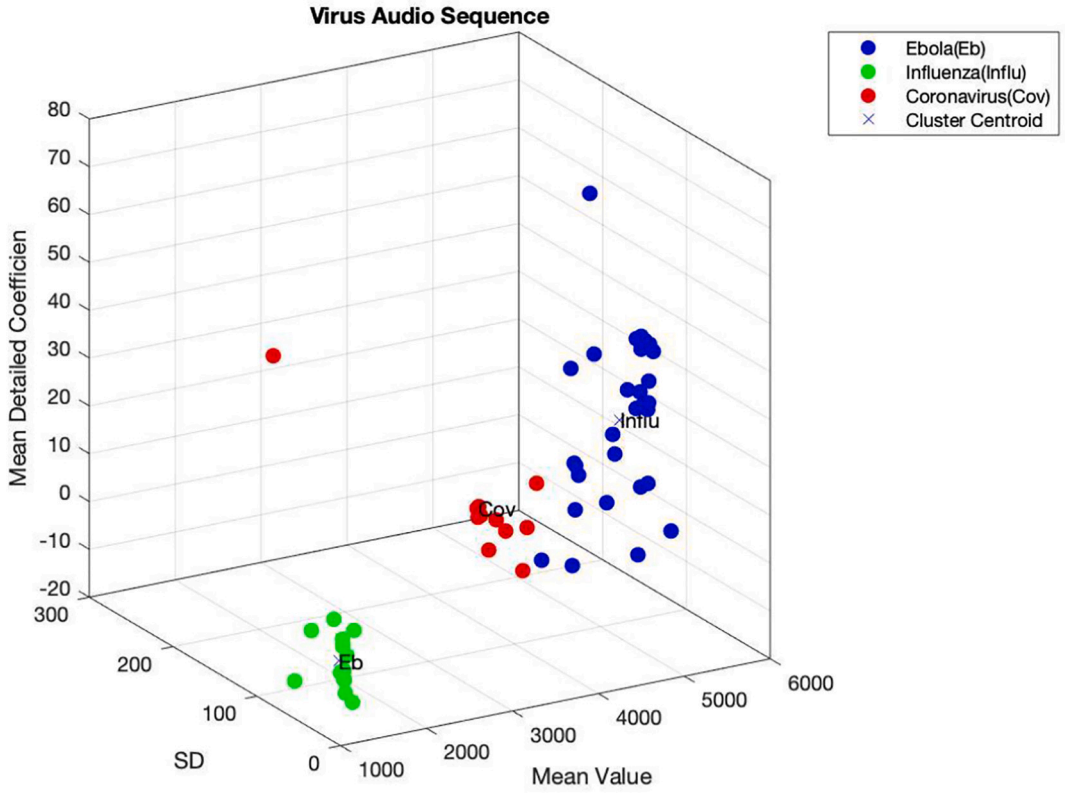


Fig. 5. Clustering of auditory sequence of the viruses.

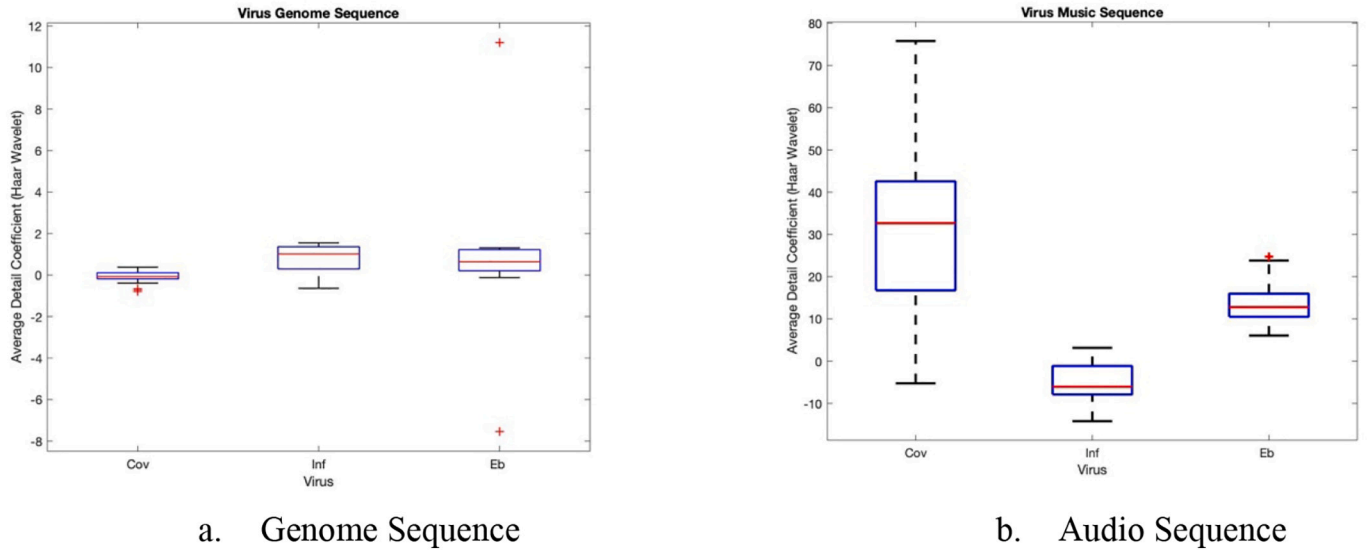


Fig. 6. Boxplot demonstrating Average detail coefficient distribution on different viruses.

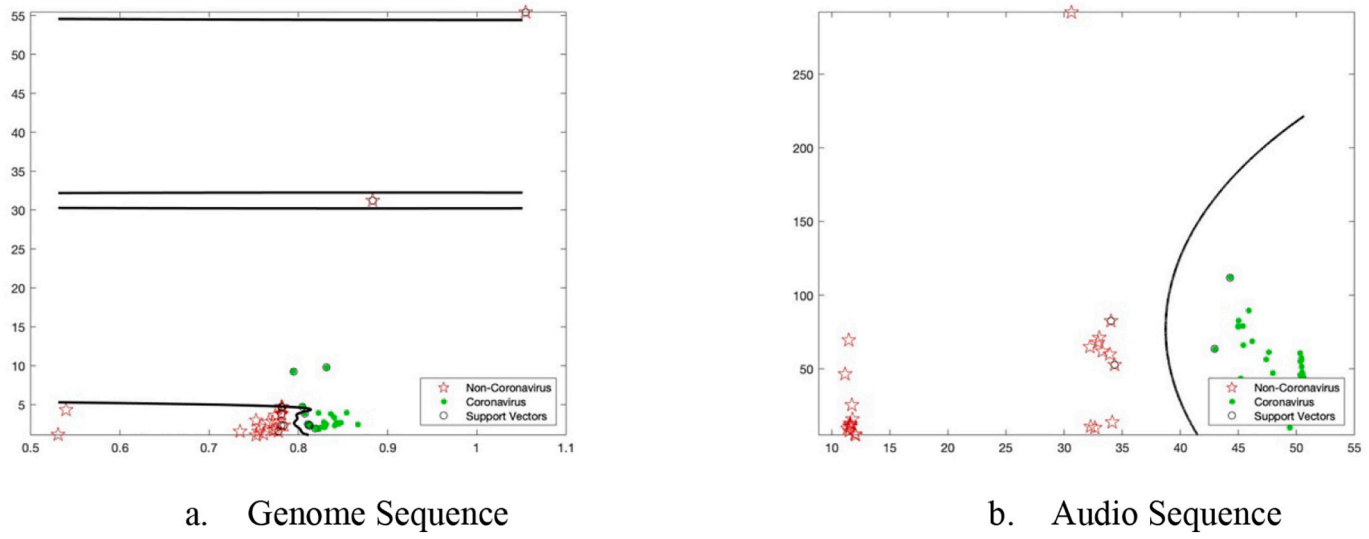


Fig. 7. Classification based on Haar wavelet coefficients.

primary protein structures are too tiny to watch, but it can be effectively heard in the music form. The virus music representation algorithm can be a beneficial tool to help in portraying the small mutation within the family (coronavirus family) in the form of music. All three scenario (among SARS-CoV, SARS-CoV-2 and MERS-CoV) show that the Euclidian distance of musical data is much higher than the protein sequence data for intra-family members. The pathogenic effect in coronavirus may enhance or limit with a small mutation which can be identified in the audio sequences of the virus.

Moreover, in the inter-family scenario, the three different types of virus (Influenza, Ebola and Coronavirus) were classified through their translated audio sequence. Therefore, the comparison shows that the more promising difference is captured in the auditory representation of the protein spikes. The proposed algorithm is computationally efficient with time complexity $O(n * k * t)$, for the 'n' length sequences, 'k' is the cluster of k-means algorithm and 't' is the number of iterations. The numerical mapping based on the physical presence of each protein and the length of the virus sequence played a dominating role towards audio translation. And the scaling factor 'log_2' made a noticeable difference in the magnitude of the audio sequence in MIDI conversion. This algorithm will be a helpful tool to find and classify virus sequences into virus family and species, and also make a difference from the other members of the same family without studying in a laboratory condition.

Declaration of Competing Interest

None.

Acknowledgment

This research work was supported by Infotech Oulu Doctoral Program.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2020.10.009.

References

[1] P. Yang, X. Wang, COVID-19: a new challenge for human beings, Cell. Mol. Immunol. (2020), https://doi.org/10.1038/s41423-020-0407-x.

[2] B. Xu, B. Gutierrez, S. Mekaru, K. Sewalk, L. Goodwin, A. Loskill, E.L. Cohn, Y. Hswen, S.C. Hill, M.M. Cobo, A.E. Zarebski, S. Li, C.H. Wu, E. Hulland, J. D. Morgan, L. Wang, K. O'Brien, S.V. Scarpino, J.S. Brownstein, O.G. Pybus, D. M. Pigott, M.U.G. Kraemer, Epidemiological data from the COVID-19 outbreak, real-time case information, Sci. Data 7 (2020) 106, https://doi.org/10.1038/s41597-020-0448-0.

[3] H.A. Rothan, S.N. Byrareddy, The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak, J. Autoimmun. (2020), https://doi.org/10.1016/j.jaut.2020.102433.

[4] Z. Sun, S. Pei, R.L. He, S.S.T. Yau, A novel numerical representation for proteins: three-dimensional chaos game representation and its extended natural vector, Comput. Struct. Biotechnol. J. 18 (2020) 1904–1913, https://doi.org/10.1016/j.csbj.2020.07.004.

[5] T. Hoang, C. Yin, S.S.T. Yau, Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison, Genomics 108 (2016) 134–142, https://doi.org/10.1016/j.ygeno.2016.08.002.

[6] T. Hoang, C. Yin, S.S.T. Yau, Splice sites detection using chaos game representation and neural network, Genomics 112 (2020) 1847–1852, https://doi.org/10.1016/j.ygeno.2019.10.018.

[7] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, Q. Zhou, Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2, Science 80 (367) (2020) 1444–1448, https://doi.org/10.1126/science.abb2762.

[8] World Health Organization, visited on (Visited on 13th June, 2020). https://www.who.int/activities/prioritizing-diseases-for-research-and-development-in-emergency-contexts, 2020.

[9] S. Xia, M. Liu, C. Wang, W. Xu, Q. Lan, S. Feng, F. Qi, L. Bao, L. Du, S. Liu, C. Qin, F. Sun, Z. Shi, Y. Zhu, S. Jiang, L. Lu, Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion, Cell Res. 2 (2020), https://doi.org/10.1038/s41422-020-0305-x.

[10] F. Wu, S. Zhao, B. Yu, Y.M. Chen, W. Wang, Z.G. Song, Y. Hu, Z.W. Tao, J.H. Tian, Y.Y. Pei, M.L. Yuan, Y.L. Zhang, F.H. Dai, Y. Liu, Q.M. Wang, J.J. Zheng, L. Xu, E. C. Holmes, Y.Z. Zhang, A new coronavirus associated with human respiratory disease in China, Nature 579 (2020) 265–269, https://doi.org/10.1038/s41586-020-2008-3.

[11] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G.F. Gao, W. Tan, A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. 382 (2020) 727–733, https://doi.org/10.1056/NEJMoa2001017.

[12] P. Zhou, X. Lou Yang, X.G. Wang, B. Hu, L. Zhang, W. Zhang, H.R. Si, Y. Zhu, B. Li, C.L. Huang, H.D. Chen, J. Chen, Y. Luo, H. Guo, R. Di Jiang, M.Q. Liu, Y. Chen, X. R. Shen, X. Wang, X.S. Zheng, K. Zhao, Q.J. Chen, F. Deng, L.L. Liu, B. Yan, F. X. Zhan, Y.Y. Wang, G.F. Xiao, Z.L. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273, https://doi.org/10.1038/s41586-020-2012-7.

[13] S. Jiang, L. Du, Z. Shi, An emerging coronavirus causing pneumonia outbreak in Wuhan, China: calling for developing therapeutic and prophylactic strategies, Emerg. Microbes Infect. 9 (2020) 275–277, https://doi.org/10.1080/22221751.2020.1723441.

[14] A.E. Gorbalenya, S.C. Baker, R.S. Baric, R.J. de Groot, C. Drosten, A.A. Gulyaeva, B. L. Haagmans, C. Lauber, A.M. Leontovich, B.W. Neuman, D. Penzar, S. Perlman, L. L.M. Poon, D.V. Samborskiy, I.A. Sidorov, I. Sola, J. Ziebuhr, The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, Nat. Microbiol. 5 (2020) 536–544, https://doi.org/10.1038/s41564-020-0695-z.

[15] O. Susumu, O. Midori, The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition, Immunogenetics 24 (1986) 71–78, https://doi.org/10.1007/BF00373112.

[16] K. Hayashi, N. Munakata, Basically musical, Nature 310 (1984) 96, https://doi.org/10.1038/310096a0.

[17] P. Gena, D. Ph, C. Strom, A Physiological Approach to DNA Music, Sixth Int. Symp. Electron. Art, 1995, pp. 83–85.

[18] P. Gena, D. Ph, C. Strom, Musical Synthesis of DNA Sequences, XI Colloq. Di Inform. Music, 1995.

[19] J. Dunn, M.A. Clark, Life music: the Sonification of proteins, Leonardo 32 (1999) 25–32, https://doi.org/10.1162/002409499552966.

[20] R. Takahashi, J.H. Miller, Conversion of amino-acid sequence in proteins to classical music: search for auditory patterns, Genome Biol. 8 (2007), https://doi.org/10.1186/gb-2007-8-5-405.

[21] R. Castagna, A. Chiolerio, V. Margaria, Music translation of tertiary protein structure: Auditory patterns of the protein folding, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2011, pp. 214–222, https://doi.org/10.1007/978-3-642-20520-0_22, 6625 LNCS.

[22] T. Paul, S. Vainio, J. Roning, Towards personalised, DNA signature derived music via the short tandem repeats (STR), in: Adv. Intell. Syst. Comput, 2019, pp. 951–964, https://doi.org/10.1007/978-3-030-01177-2_69.

[23] B.C. Florea, MIDI-based controller of electrical drives, in: Proc. 2014 6th Int. Conf. Electron. Comput. Artif. Intell. ECAI 2014, 2015, pp. 27–30, https://doi.org/10.1109/ECAI.2014.7090159.

[24] R.P. Bywater, J.N. Middleton, Melody discrimination and protein fold classification, Heliyon 2 (2016), https://doi.org/10.1016/j.heliyon.2016.e00175.

[25] A.K. Belman, T. Paul, L. Wang, S.S. Iyengar, P. Śniatała, Z. Jin, V.V. Phoha, S. Vainio, J. Roning, Authentication by mapping keystrokes to music: the melody of typing, Int. Conf. Artif. Intell. Signal Process. AISP 2020 (2020), https://doi.org/10.1109/AISP48273.2020.9073125.

[26] Massachusetts Institute of Technology, Visited on 13th June, http://news.mit.edu/2020/qa-markus-buehler-setting-coronavirus-and-ai-inspired-proteins-to-music-0402, 2020.

[27] M. Marques, V. Oliveira, S. Vieira, A.C. Rosa, Music composition using genetic evolutionary algorithms, in: Proc. 2000 Congr. Evol. Comput. CEC 2000. 1, 2000, pp. 714–719, https://doi.org/10.1109/CEC.2000.870368.

[28] F. Bertino, C. Chuan, J. Peroune, The Musical Gene: Generating Harmonic Patterns from Sequenced DNA of E. coli Bacteria to Compose Music, Work, Visited on 13th June, 2020.

[29] M.I.D.I. Ken Schutte, Matlab Toolbox, Visited on 13th June, https://github.com/kts/matlab-midi, 2020.

[30] D. Wei, Q. Jiang, Y. Wei, S. Wang, A novel hierarchical clustering algorithm for gene sequences, BMC Bioinforma. 13 (1) (2012), https://doi.org/10.1186/1471-2105-13-174.

[31] R. Dong, L. He, R.L. He, S.S.T. Yau, A novel approach to clustering genome sequences using inter-nucleotide covariance, Front. Pharmacol. 10 (2019), https://doi.org/10.3389/fgene.2019.00234.

[32] M. Deng, C. Yu, Q. Liang, R.L. He, S.S.T. Yau, A novel method of characterizing genetic sequences: genome space with biological distance and applications, PLoS One 6 (2011), https://doi.org/10.1371/journal.pone.0017293.

[33] M. Steinegger, J. Söding, Clustering huge protein sequence sets in linear time, Nat. Commun. 9 (2018), https://doi.org/10.1038/s41467-018-04964-5.

[34] B.T. James, B.B. Luczak, H.Z. Girgis, MeShClust: an intelligent tool for clustering DNA sequences, Nucleic Acids Res. 46 (2018) e83, https://doi.org/10.1093/nar/gky315.

[35] J. Lin, J. Wei, D. Adjeroh, B.H. Jiang, Y. Jiang, SSAW: a new sequence similarity analysis method based on the stationary discrete wavelet transform, BMC Bioinforma. 19 (2018) 1–11, https://doi.org/10.1186/s12859-018-2155-9.

[36] D.W. Liu, R.P. Jia, C.F. Wang, N. Arunkumar, K. Narasimhan, M. Udayakumar, V. Elamaran, Automated detection of cancerous genomic sequences using genomic signal processing and machine learning, Futur. Gener. Comput. Syst. 98 (2019) 233–237, https://doi.org/10.1016/j.future.2018.12.041.

[37] T. Paul, S. Vainio, J. Roning, Haar wavelet based approach for Short Tandem Repeats(STR) Detection, in: 2019 IEEE 19th Int. Symp. Signal Process. Inf. Technol. ISSPIT 2019, 2019, pp. 1–6, https://doi.org/10.1109/ISSPIT47144.2019.9001825.

[38] R.B.A. Bakar, J. Watada, W. Pedrycz, DNA approach to solve clustering problem based on a mutual order, BioSystems 91 (2008) 1–12, https://doi.org/10.1016/j.biosystems.2007.06.002.

[39] B. Kenidra, M. Benmohammed, A. Beghriche, Z. Benmounah, A partitional approach for genomic-data clustering combined with K-Means algorithm, in: Proc. - 19th IEEE Int. Conf. Comput. Sci. Eng. 14th IEEE Int. Conf. Embed. Ubiquitous Comput. 15th Int. Symp. Distrib. Comput. Appl. to Business, Engi, 2017, pp. 114–121, https://doi.org/10.1109/CSE-EUC-DCABES.2016.170.

[40] T.K. Seo, Classification of nucleotide sequences using support vector machines, J. Mol. Evol. 71 (2010) 250–267, https://doi.org/10.1007/s00239-010-9380-9.

[41] T. Wang, M. Herbster, I.S. Mian, Virus Genome Sequence Classification using Features based on Nucleotides, Words and Compression, 2018, pp. 1–36. http://arxiv.org/abs/1809.03950.

[42] C. Zou, J. Gong, H. Li, An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis, BMC Bioinforma. 14 (2013), https://doi.org/10.1186/1471-2105-14-90.

[43] NCBI Database, Visited on 13th June, https://www.ncbi.nlm.nih.gov/, 2020.

[44] C. Drosten, S. Günther, W. Preiser, S. Van der Werf, H.R. Brodt, S. Becker, H. Rabenau, M. Panning, L. Kolesnikova, R.A.M. Fouchier, A. Berger, A. M. Burguière, J. Cinatl, M. Eickmann, N. Escriou, K. Grywna, S. Kramme, J. C. Manuguerra, S. Müller, V. Rickerts, M. Stürmer, S. Vieth, H.D. Klenk, A.D.M. E. Osterhaus, H. Schmitz, H.W. Doerr, Identification of a novel coronavirus in patients with severe acute respiratory syndrome, N. Engl. J. Med. 348 (2003) 1967–1976, https://doi.org/10.1056/NEJMoa030747.

[45] R. Lu, Y. Wang, W. Wang, K. Nie, Y. Zhao, J. Su, Y. Deng, W. Zhou, Y. Li, H. Wang, W. Wang, C. Ke, X. Ma, G. Wu, W. Tan, Complete genome sequence of Middle East respiratory syndrome coronavirus (MERS-CoV) from the first imported MERS-CoV case in China, Genome Announc. 3 (2015) 2014–2015, https://doi.org/10.1128/genomeA.00818-15.

[46] G.N. Kouziokas, SVM kernel based on particle swarm optimized vector and Bayesian optimized SVM in atmospheric particulate matter forecasting, Appl. Soft Comput. J. 93 (2020), 106410, https://doi.org/10.1016/j.asoc.2020.106410.

[47] R. de Groot, S. Baker, R. Baric, L. Enjuanes, A. Gorbalenya, K. Holmes, S. Perlman, L. Poon, P. Rottier, P. Talbot, P. Woo, J. Ziebuhr, Part II – The Positive Sense Single Stranded RNA Viruses Family Coronaviridae, Virus Taxon. Ninth Rep. Int. Comm. Taxon. Viruses, 2012, pp. 806–828, https://doi.org/10.1016/B978-0-12-384684-6.00068-9.

[48] S. Jones, R. Prasad, A.S. Nair, S. Dharmaseelan, R. Usha, R.R. Nair, R.M. Pillai, Whole-Genome Sequences of Influenza A(H1N1)pdm09 Virus Isolates from Kerala, India 5, 2015, pp. 9–10, https://doi.org/10.1128/genomeA.00598-17.

[49] NCBI, Database Ebolavirus, Visited on 13th June, https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=186536, 2020.