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 112 (2020) 4701–4714

Contents lists available at ScienceDirect

Genomics

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

A new graph-theoretic approach to determine the similarity of genome sequences based on nucleotide triplets

Subhram Das[a,*], Arijit Das[a], D.K. Bhattacharya[b], D.N. Tibarewala[c]

[a] Computer Science and Engineering, Narula Institute of Technology, Kolkata 700109, India
[b] Pure Mathematics, Calcutta University, Kolkata 700019, India
[c] Bio-Science and Engineering, Jadavpur University, Kolkata 700032, India.

ARTICLE INFO

Keywords:
Sequence comparison
Evolutionary relationship
Alignment-based method
Alignment-free method
Bipartite graph
Phylogenetic tree

ABSTRACT

Methods of finding sequence similarity play a significant role in computational biology. Owing to the rapid increase of genome sequences in public databases, the evolutionary relationship of species becomes more challenging. But traditional alignment-based methods are found inappropriate due to their time-consuming nature. Therefore, it is necessary to find a faster method, which applies to species phylogeny. In this paper, a new graph-theory based alignment-free sequence comparison method is proposed. A complete-bipartite graph is used to represent each genome sequence based on its nucleotide triplets. Subsequently, with the help of the weights of edges of the graph, a vector descriptor is formed. Finally, the phylogenetic tree is drawn using the UPGMA algorithm. In the present case, the datasets for comparison are related to mammals, viruses, and bacteria. In most of the cases, the phylogeny in the present case is found to be more satisfactory as compared to earlier methods.

1. Introduction

Biological sequence comparison is an emerging research area for the scientist in the field of Genetics. The rapid increase of information in genome databases inspired researchers to develop methods for analyzing this information. Comparative studies of biological sequences, their proper clustering, and their evolutionary relationship are some important challenges. As a very large number of genome sequences with unequal lengths appears in different species and as we are aiming at low time complexity, so it becomes necessary to find new approaches. In this connection, traditional alignment-based methods are the earlier ones. But they are treated as outdated approaches because of their limitations [1,2]. For instance, the multiple-alignment method, one of the alignment-based methods for DNA sequence similarity, possesses time complexity $O(n^2)$, where n is the length of DNA sequence. It clearly shows very high time complexity [3]. The alternatives to alignment-based methods are alignment-free methods [4–7]. These consist of numerical approaches along with graphical representation. There are several attempts of different authors to compare genome sequences using different alignment-free methods. For pictorial inspection and clustering between genomic sequences, the graphical representation method is one of the most efficient ones. For such representation, first of all, the data set is replaced by numerical values, and graphs are drawn accordingly. Such graphs in two-dimension [8–10], three-dimension [11–16], and even four-dimension [17–19] are obtained from time to time. Next from the graphs, some descriptors are obtained, which are used for obtaining similarity/dissimilarity matrices for sequence comparison [20]. There are two ways of finding descriptors. One is called graphical descriptors [21]. They are obtained directly from the data points of the graph. The second one is called matrix form [22] of descriptor. In this case, some matrices are obtained from the data points of the graph. Now the descriptors correspond to some real values associated with these matrices in the form of its maximum eigenvalue or the row-max. All such methods are basically the same; they differ only in the choice of the descriptors and the choice of distance measures. Now we mention methods, which are by and large probabilistic in nature [23–25]. One such method [23] is to be mentioned especially because first of all a 2D representation of DNA sequences are obtained from the data points. Next, a probability vector is obtained, which is taken as the descriptor. Thus the distance measure is

* Corresponding author.
E-mail address: subhram.das@nit.ac.in (S. Das).

https://doi.org/10.1016/j.ygeno.2020.08.023
Received 9 September 2019; Received in revised form 15 July 2020; Accepted 17 August 2020
Available online 19 August 2020

a probabilistic one. In fact, it is taken as the symmetric form of Kullback-Leibler divergence. Next, we mention the general probabilistic approach to DNA sequence comparison. These methods are generally called composition vector (CV) methods. They are based on either k-mer or word frequency [26–28]. Using such methods, genome sequence comparison is done by calculating the frequency of substring k, where k is any positive number. It is seen that for the choice of some values of k and choice of some particular distance measure, the results of the comparison are good for some sequences, but they are not so in other cases. Therefore it becomes necessary to find some optimal value of k and some proper choice of distance measure, under which better results are obtained in most of the cases. Recently it is shown in [29] that when k is taken to be 3 and when the information-based similarity index is taken as the distance measure, the k-mer method is workable in all most all cases. Actually, plenty of research work has already been done based on this k-mer method for genome sequence analysis. Feature frequency profile (FFP) [30] method is another such example, where a vector descriptor for a genome sequence is formed with the help of frequency of the substring of length k. The general name of such methods is the Complete Composition Vector (CCV) [31] methods. CCV methods are further modified to optimize the CV [32] method. Finally Improved Complete Composition Vector (ICCV) method [33] is obtained by optimizing both CV and CCV methods. ICCV method is found to be more robust and efficient. Now we mention the fuzzy integral similarity method [34], which is also under the classification of k-mer or word frequency method. In this method, the Markov chain is used as an estimated parameter to calculate similarity scores between two DNA sequences. Lastly, we mention another approach to DNA sequence comparison, which is graph-theoretic. Although it is less complicated but very much useful, still it is rarely used in genome sequence comparison. Most of these methods are weighted directed multi-graphs. In [35], the authors form a 4 × 4 similarity matrix based on the weights between 16 di-nucleotide pair. Similar attempts are also taken up in [36], where they use weighted directed multi-graphs to obtain the phylogeny of a family of species. In this paper, DNA sequences are considered as a combination of triplets of nucleotide and are represented in the form of a complete bipartite graph. The notion of this bipartite graph is very interesting in graph theory. It has the advantage of using a pair of a vertex set and their inter-related weights to find a distance matrix. The present work applies the use of such a bipartite graph in the DNA sequence comparison for the first time and it also significantly differs from other graph-theoretical methods. We verify our proposed approach with the benchmark datasets of the following species. These are 41 Mammalian mitochondrial genomes, 30 coronaviruses, and 4 non-corona virus genomes, 48 Hepatitis E viruses (HEV), 53 complete genome sequences of Tomato yellow leaf curl viruses (TYLCV), 59 bacterial genomes, 59 Ebola viruses, and 38 Influenza A Virus. Further, we compare our results with those obtained by other advanced sequence analysis techniques on the same data set. To check the efficiency of our method, we also calculate its time-complexity. The proposed graph-theoretic approach helps to find evolutionary relationships efficiently amongst the species without any genetic involvement.

2. Methods

The graph-theoretic approach is one of the convenient ways of analyzing genome sequences. In our work, we use a complete bipartite graph to represent genome sequences. A complete graph is one where any two vertices are connected by an edge. In particular, a bipartite graph has some special features. Such a graph has two independent sets of graph vertices and that no two graph vertices of the same set are adjacent. A bipartite graph [37] G is denoted by the pair (V, K), where $V = (V_1, V_2)$ are the two sets of vertices and K represents the edges of the graph.

Fig. 1 shows that it is a complete bipartite graph because each of the vertices A, B, C from the first set V_1 are connected with each of the vertex D, E, F of the second set V_2 and that neither A, B, C nor D, E, F are adjacent to each other. In this paper, we have chosen a complete bipartite graph to represent genome sequences because nucleotide triplets can easily be expressed by nodes of the graph and the edges of the graph are used to calculate vector descriptor.

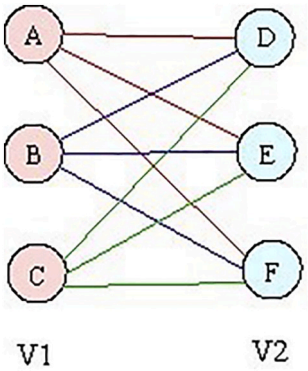

Fig. 1. Example of a complete bipartite graph.

2.1. Construction of bipartite graph using nucleotide triplets

As shown in Fig. 2, we consider two independent sets V1 and V2, where V1 consists of (A, C, T, G), the four nucleotide bases as vertices and V2 consists of (AA, AC, AT, AG, CA, CC, CT, CG, TA, TC, TT, TG, GA, GC, GT, GG), the sixteen di-nucleotides as vertices. All the vertices of set V_1 are connected with every vertex of set V_2. This way it becomes a complete bipartite graph.

2.2. Calculation of weighted vector

First of all, n-1 nucleotide triplets are considered from the DNA sequence of length n, in an overlapping manner. Then each nucleotide triplet is thought of as a combination of a nucleotide and a di-nucleotide. In this fashion, 64 nucleotide triplets are represented with the help of a bipartite graph. For the sake of calculation, we assume that the weight of each edge is one. With this assumption, we finally calculate the weighted vectors of 64 components from the given sequences.

2.3. Computation of distance matrix

Let $S_1$and S_2 be two DNA sequences of two different species. Let x_i and y_i be the corresponding vector values of the sequence S_1 and S_2, where $i = 1, 2, 3, \ldots, 64$. Now the distance between two sequences S_1 and S_2 is calculated using the formula $D(S_1, S_2) = \frac{\sum_1^{64} | x_i - y_i |}{64} D(S_1, S_2) = \frac{\sum_1^{64} | x_i - y_i |}{64}$. Thus we get the similarity matrix for a set of DNA sequences S_1, S_2, S_3, S_4…. , S_m, where m is the number of sequences.

In our work, the distance is calculated between two 4 × 16 matrices. In mathematics, such distance is measured between two 64 dimensional vectors, because a 4 × 16 real matrix is homeomorphic to a 64-dimensional vector. So any metric on R^{64} is sufficient for the purpose. Naturally Manhattan metric is one such choice. Hence $d(X, Y) = \Sigma_{i=1}{}^{64} |x_i - y_i|$ may be a suitable one. But in the present case, x_i and y_i are not arbitrary real numbers. Rather they are the weighted components of the weighted matrix of a graph. Naturally, they are interrelated. In such a case the average of the Manhattan metric which is also a metric is found to be very much useful in matrix comparison. This has the standard name weighted distance(WD) and in this case, it is given by $d(X, Y) = \frac{1}{64} \sum_{i=1}^{64} |x_i - y_i|$.

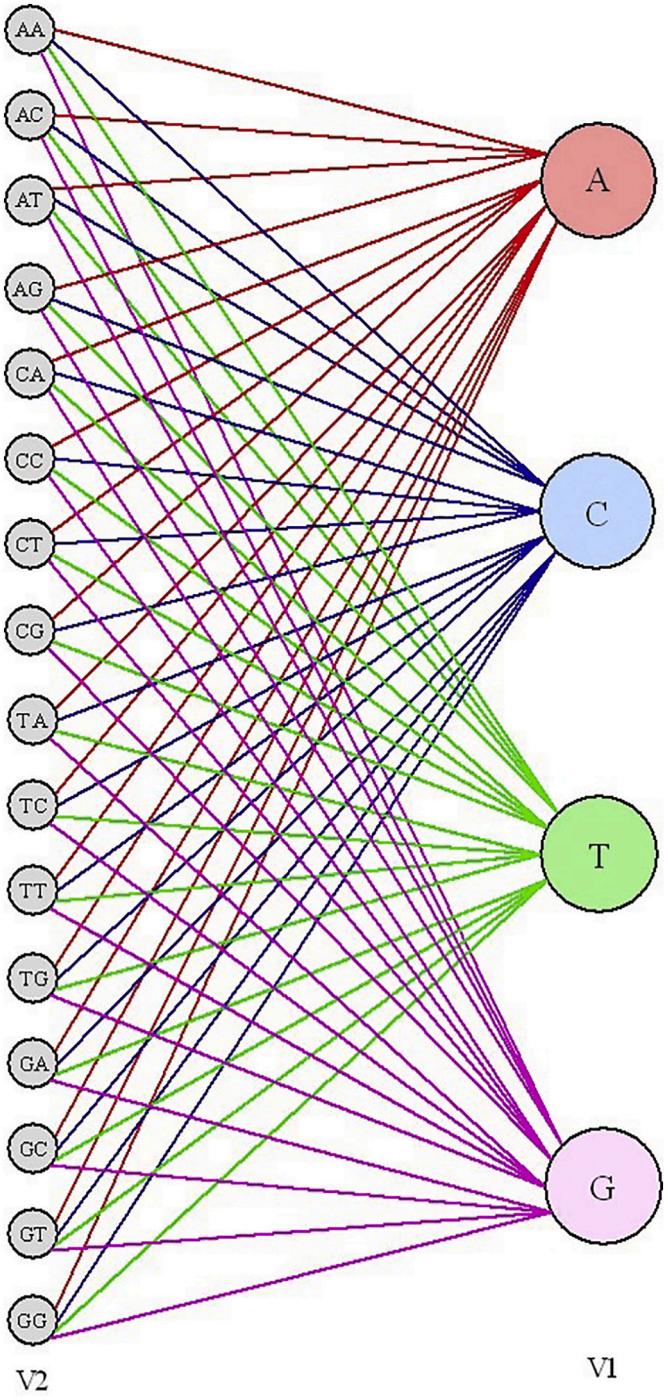

Fig. 2. Representation of nucleotide triplets using a complete bipartite graph.

2.4. Construction of phylogenetic tree

A standard UPGMA algorithm is applied to construct a phylogenetic tree using MEGA7 software [38].

Fig. 3 describes the flow-chart for our proposed graph-theoretic approach.

Table 1 describes the algorithm of our proposed graph-theoretic approach.

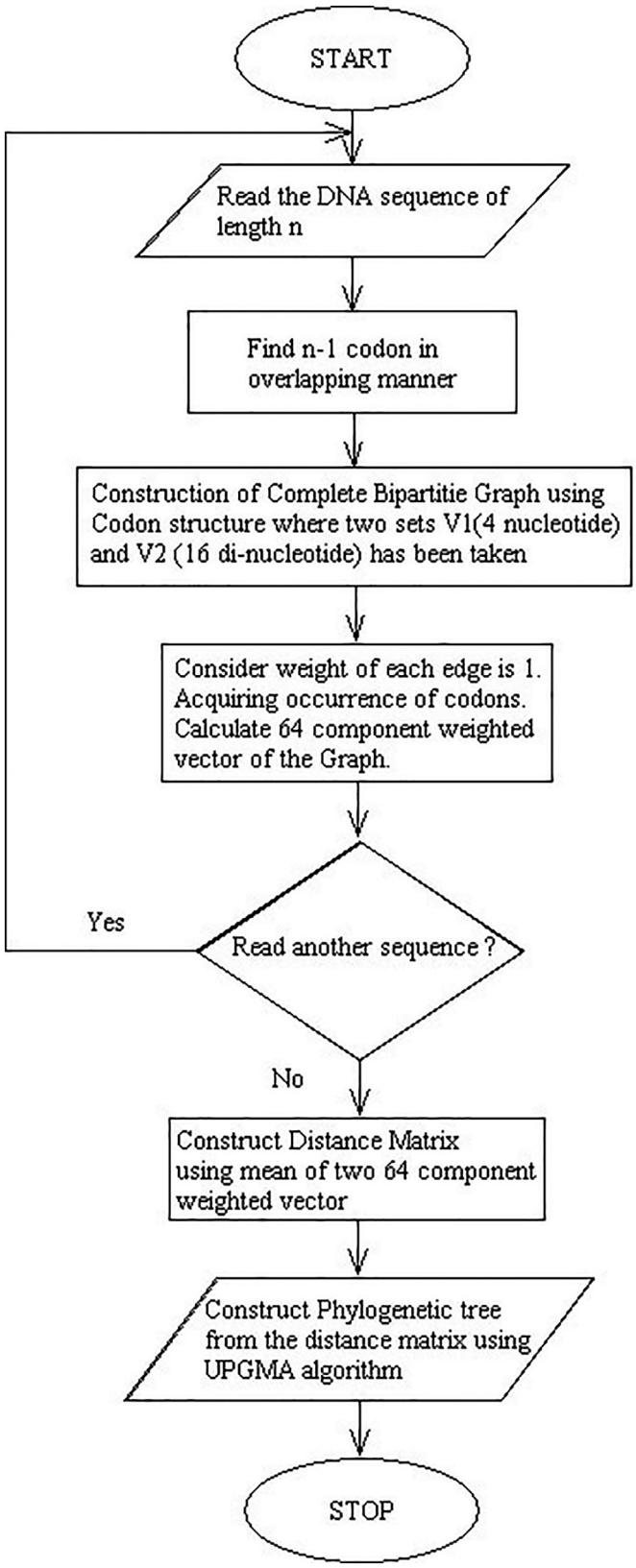

Fig. 3. Flow-chart of our proposed graph-theoretic method.

2.5. Time-complexity

To find the efficiency of the algorithm, we examine the time complexity of our proposed algorithm. We assume that all the operations performed inside the algorithm consume the same unit of time. Now, we divide the whole computational process into three stages.

In the first stage, frequencies of nucleotide triplets are calculated. For each DNA sequence of length n, we are to search the *n-1* number of

Table 1
Algorithm of our proposed Graph theoretic approach.

```
1. BASES[ ] = ["A","C","T","G"]
2. STRINGS[ ] = all possible combination of strings of length 2 using characters in BASES[ ]
3. FILES[ ] = name of all input files
4. NO_OF_FILES = size_of(FILES)
5. initialize DATA[ ][ ][ ] with 0
6. initialize DISTANCE_MATRIX[ ][ ] with 0
7. for each file f in the list FILES[ ]do
        7.1 Read the DNA sequence D from the file f
        7.2 Remove all other alphabets and symbols from D except the characters of the list BASES[ ]
        7.3 for each pattern p in the list STRINGS[ ] do
                7.3.1 for each character x in the list BASES[ ] do
                        7.3.1.1 DATA[f][p][x] = no of pattern p+xin the file f
                7.3.2 end for
        7.4 end for
8. end for
9. fori = 0 to ( NO_OF_FILES – 1) do
        9.1 f1 = FILES[i]
        9.2 for each file f2 in the range FILES[i+1] to FILES[NO_OF_FILES]do
                9.2.1 sum = 0
                9.2.2 for each pattern p in the list STRINGS[] do
                        9.2.2.1 for each character x in the list BASES[ ] do
                                9.2.2.1.1 sum = sum + mod of(DATA[f1][p][x] – DATA[f2][p][x] )
                        9.2.2.2 end for
                9.2.3 end for
                9.2.4 DISTANCE_MATRIX[f1][f2] = sum/64
        9.3 end for
10. end for
11. return DISTANCE_MATRIX[ ][ ]
12.using DISTANCE_MATRIX phylogenetic tree is constructed using MEGA7 software
```

such triplets. Therefore, the time complexity is O(n).

Number of sequences	No of sequence to compare
1	k-1
2	k-2
3	k-3
.....	.....
k-1	1

In the second stage, to represent a 64-component vector from the n-1 number of nucleotide triplets, each DNA sequence takes time complexity O(64*n) or O(n). Now, we consider k number of different sequences for comparison. So, the total time complexity in stage two is O ($k \times n$).

For the third or final stage, we calculate the distance matrix between k no of sequences. So total no of complexity for comparisons = O ($[k \times (k - 1)] / 2$) = O($k \times k$). Now for each comparison, it takes O (n), so time complexity of stage three is O($k^2 \times n$). Therefore, the total time complexity is O($k \times n$) + O($k^2 \times n$) = O($k^2 \times n$). Thus it is clear that for calculating the distance between a pair of sequences of equal length n, the time complexity depends only on the value of n.

2.6. 2.6 Simplicity

Time complexity is a major criterion to claim a method to be efficient. It is good that our method is less time-consuming than other methods. But overall simplicity is another very important feature to prefer one method over the other. The other methods are not so simple as the present one. We originate a 64-component vector based on the complete-bipartite graph to compare genome sequences. Our approach produces only 64-component vectors for each genome sequence without applying any prior alignment techniques. To execute this method, no additional parameter is required. The novelty of this approach is that only combinations of nucleotide triplets are considered to obtain such a complete-bipartite graph. To obtain the vertices and the sides of the graph requires simple steps and to write the weighted matrix of the graph and to calculate the distances, involve simple procedures. Therefore, this method is very simple to execute.

3. Results and discussion

To validate our method, we use the dataset of mitochondrial genomes of 41 mammalian, 30 coronavirus, and 4 non-corona virus genomes, 48 Hepatitis E viruses (HEV), 53 complete genome sequences of Tomato yellow leaf curl viruses (TYLCV), 59 different bacterial genomes, 59 Ebola viruses, and 38 Influenza A Virus. We use MEGA7 to construct a phylogenetic tree for each dataset. Our constructed phylogenetic trees are then compared with the results obtained from some other advanced genome sequence comparison methods like Feature Frequency Profiles (FFP) method [30], Fuzzy Integral Similarity method [34], ClustalW method, Multiple Encoding Vector method [40], Fast Vector method [40], Weighted Measure method [42], Probabilistic method [23], K-mer method [29]. From the results, we can see the efficiency in terms of time complexity and accuracy of our new method is better than or equal to all other previously published methods on the same datasets. The following are the detailed discussion of our method with different datasets.

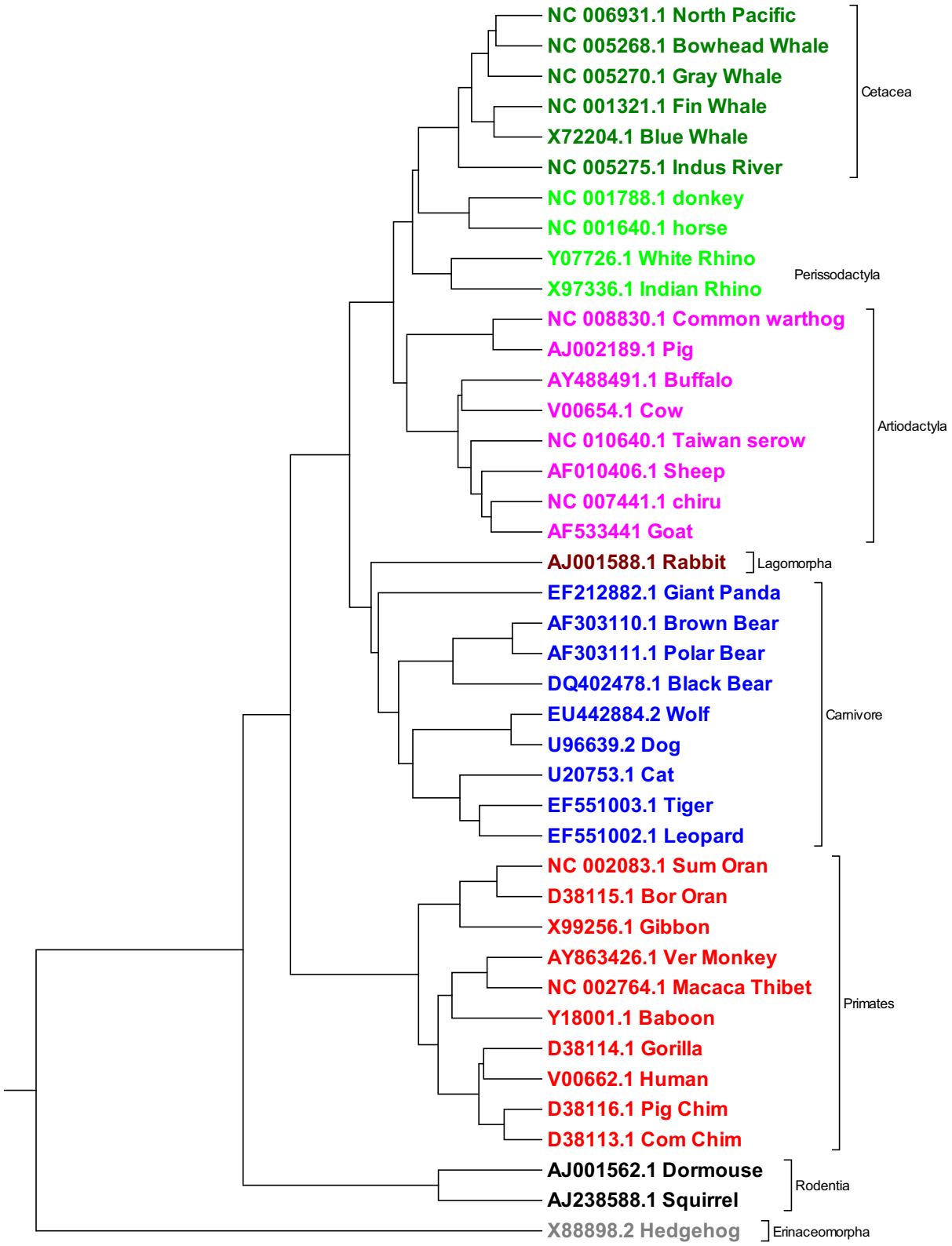

Fig. 4. Phylogenetic tree of 41 mitochondrial genomes under our graph-theoretic approach.

3.1. Analysis of 41 mammalian mitochondrial genomes using Phylogenetic tree

To verify our proposed method, we choose a dataset consisting of complete mitochondrial genomes (mtDNA) [39] of 41 mammalians. For each pair, the average bases are 16,500. The structure of mtDNA is circular and double-stranded for all the mammals. One of the strands is called a heavy strand; it is rich in guanine. The other strand is called the light strand; it is rich in cytosine. For our present experiment, we choose the heavy strands only which are highly conserved possessing a

fast mutation rate.

In our graph-theoretical method, it gives a correct classification of 41 species into eight clusters: Primates (red), Carnivore (blue), Cetacea (green), Perissodactyla (light green), Artiodactyla (pink), Lagomorpha (maroon), Rodentia (black) and Erinaceomorpha (grey). We now compare our tree (Fig. 4), with phylogenetic trees (Supplementary Figs. S1, S2, S3, and S4) constructed by other methods [30,34,40] shown in the supplementary file. According to the Feature frequency profiles (FFP) [30] method using k = 7 shown in Fig. S1 (in the supplementary file), it is seen that most of the classifications are not proper. The cluster Perissodactyla contains four species. But here they are distributed into two clusters. Indus River Dolphin is scheduled to belong to Cetacea. But it is detached from other species of Cetacea. Cetacea is also divided into two groups. The species belonging to the clusters Primates, Artiodactyla and Carnivore are all erroneously separated into different clusters. Overall in FFP methods, all eight types of species are not classified properly. The phylogenetic tree constructed according to the Fuzzy integral similarity method [34] is shown in Supplementary Fig. S2. It gives correct classification for only five clusters. Both Rabbit and Cat are misclassified. Rabbit is classified with Carnivore (blue) and Cat is placed in the cluster of Artiodactyla (pink). As a result, three clusters are not properly formed. Erinaceomorpha (grey) is also clustered with Rodentia (black), which should be separated from all other species. Therefore, it is seen that the classification is not proper. The Phylogenetic tree constructed with the help of multiple encoding vector methods [40] as shown in Fig. S3 (in the supplementary file) gives correct classification for all eight clusters. But Vervet Monkey and Macaca-Thibetana of family Primates are not grouped into a single clade under the subgroup of Cercopithecidae. We also compare our phylogenetic tree with that constructed under the ClustalW method given in supplementary Fig. S4. Here Vervet Monkey and Macaca-Thibetana are not visible in a single clade in the same sister cluster. Thus phylogenetic tree as shown in Fig. 4, shows better clustering as compared to those given in Figs. S1, S2, S3, and S4. Therefore, the phylogenetic tree of our method provides a better result than ClustalW and multiple encoding vector methods.

3.2. Analysis of 30 coronaviruses and 4 non-corona virus genomes using phylogenetic tree

Coronaviruses form a subfamily of Coronaviridae. The number of nucleotides in their genomes varies from 25,000 to 32,000. Some subtypes of coronaviruses are of major concern, as they can infect humans and cause severe respiratory and gastrointestinal problems. Middle East respiratory syndrome coronavirus (MERS-CoV) is one such type of coronavirus. It is at the root of the recent respiratory problem causing a high fatality rate. Hence it became necessary to classify and find an evolutionary relationship between one of SARS types of viruses, called pandemics. Now such a phylogenetic tree on 30 coronaviruses and 4 non-corona viruses (Supplementary Table S2) is constructed by our graph-theoretic approach using the UPGMA algorithm. The corresponding phylogenetic tree is given in Fig. 5 .

Fig. 5 clearly shows proper classifications of all the viruses in six distinct clusters according to their host types. Now we compare our phylogenetic tree with those given in Figs. S5, S6, S7, and S8 (in the supplementary file) obtained by other methods [30,34,40]. Phylogenetic tree generated by the Fuzzy integral similarity method [34] as given in Fig. S5 is unable to cluster group 1. HCoV-NL63 of group 1 is clustered with HCoV-HKU1 of group 5. Also, HCoV-229E and PEDV of group 1 are wrongly clustered with group 2. Now we consider Fig. S6 constructed by the FFP method [30] under substrings of length six and Fig. S7 generated by the ClustalW method. In both cases, the four non-corona viruses are not clustered collectively. But interestingly the phylogenetic tree obtained by the Fast vector method [40] shown in Fig, S8 is consistent with that of our present paper. Our result is also found consistent, with the result found in [41].

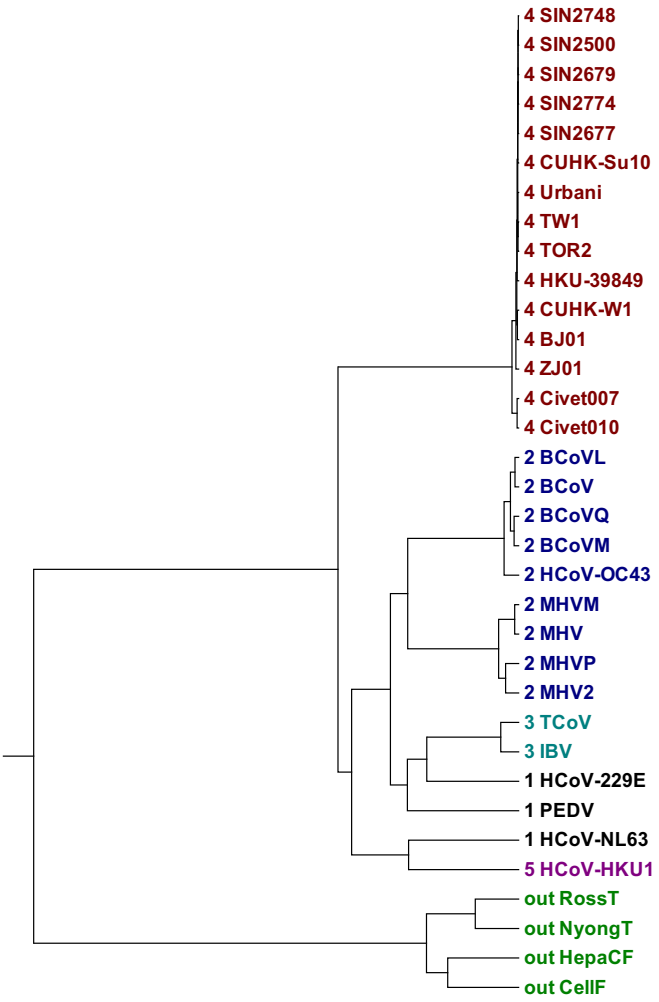

Fig. 5. Phylogenetic tree of five different types of 30 coronaviruses and 4 non-corona viruses based on our graph-theoretic approach.

3.3. Analysis of 48 Hepatitis E viruses (HEV) using phylogenetic tree

The characteristics of this type of viruses are that they are non-enveloped and single-stranded RNA. The number of nucleotides in each genome is approximately 7200. Out of all viruses of types A, B, C, D, and E, the hepatitis-E virus is the only animal-host virus. It is more responsible for the acute condition of the disease. So it is important to study relations between HEV sequences. In the earlier research, many workers have obtained the phylogenetic tree of HEV. To analyze the phylogenetic relationship, the whole genome sequences of 48 HEV are chosen in our present work dealing with the graph-theoretic approach (Supplementary Table S3). A corresponding phylogenetic tree is shown in Fig. 6.

It clearly shows proper classifications of all the 48 HEV in four distinct genotypes. In fact, genotype I consist of 16 HEV strains. These are B1(Burma), B2(Burma), I1(India), I2(India), I3(India), Yam-67(India), C1(China), C2(China), C3(China), C4(China), ChinaHeiBei (China), P1(Pakistan), P2(Pakistan), NP1(Nepal), Morocco (Morocco), and T3 (Chad). Genotype II contains only 1 HEV strain, M1(Mexico).

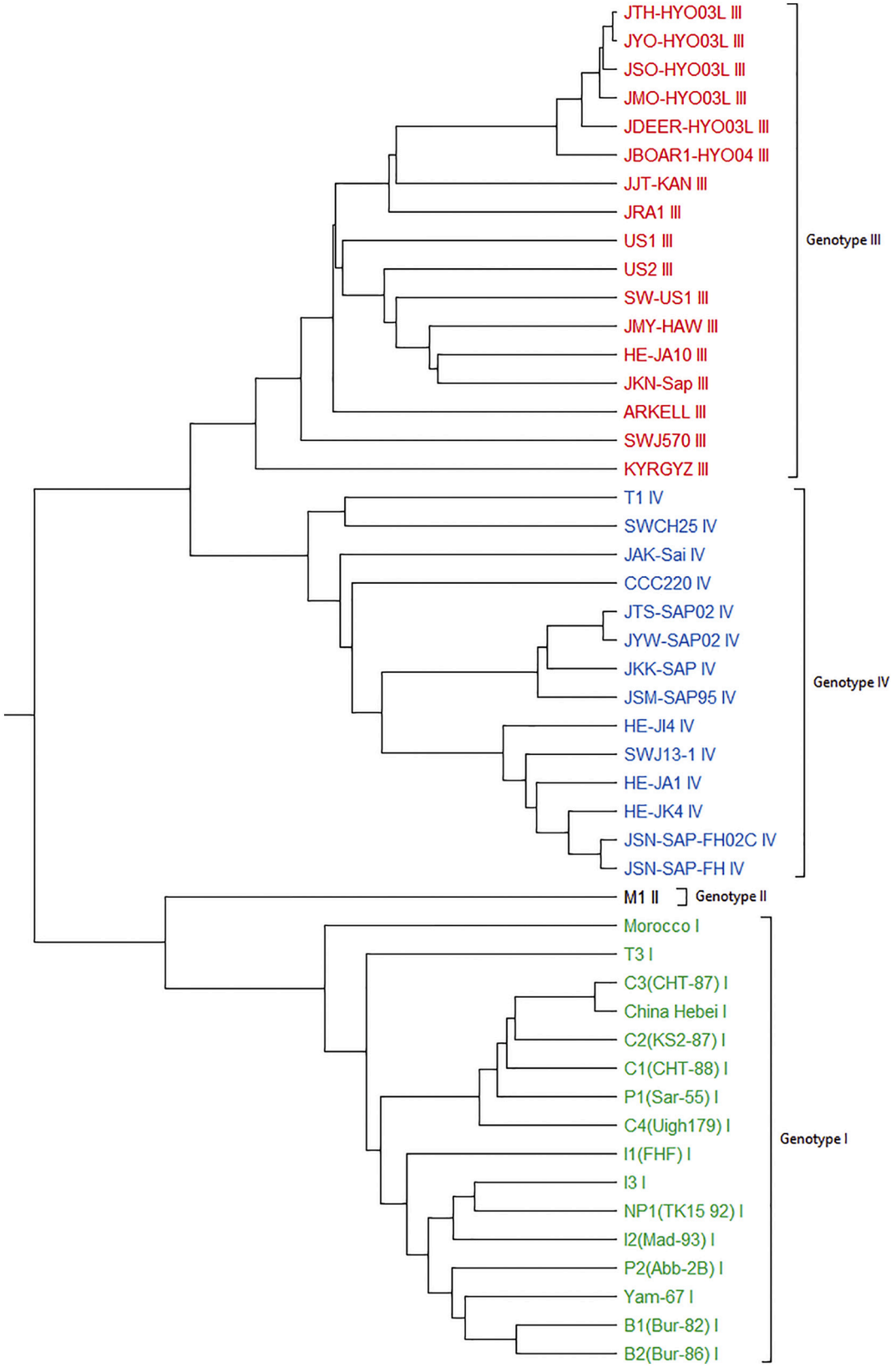

Fig. 6. The phylogenetic tree of 48 Hepatitis E virus (HEV) whole genomes constructed by our graph-theoretic approach.

Genotype III consists of 14 HEV strains, CCC220 (China), JAK-Sai (Japan), HE-JI4 (Japan), SWJ13–1 (China), HE-JA1b (Japan), HE-JK4 (Japan), JSN-SAP-FH (Japan), JSN-SAPFH020C (Japan), JSM-SAP95 (Japan), JKK-SAP (Japan), JTS-SAP02 (Japan), JYW-SAP02 (Japan), SWCH25 (China) and T1 (USA). Finally, Genotype IV includes 17 HEV strains. They are KYRGYZ (Kyrgyzstan). SWJ570 (Japan), JJT-KAN (Japan), JRA1 (Japan), JBOAR1-HY004 (Japan), JSO-HY003L (Japan), JDEER-HY003L (Japan), JMO-HY003L (Japan), JTH-HY003L (Japan),

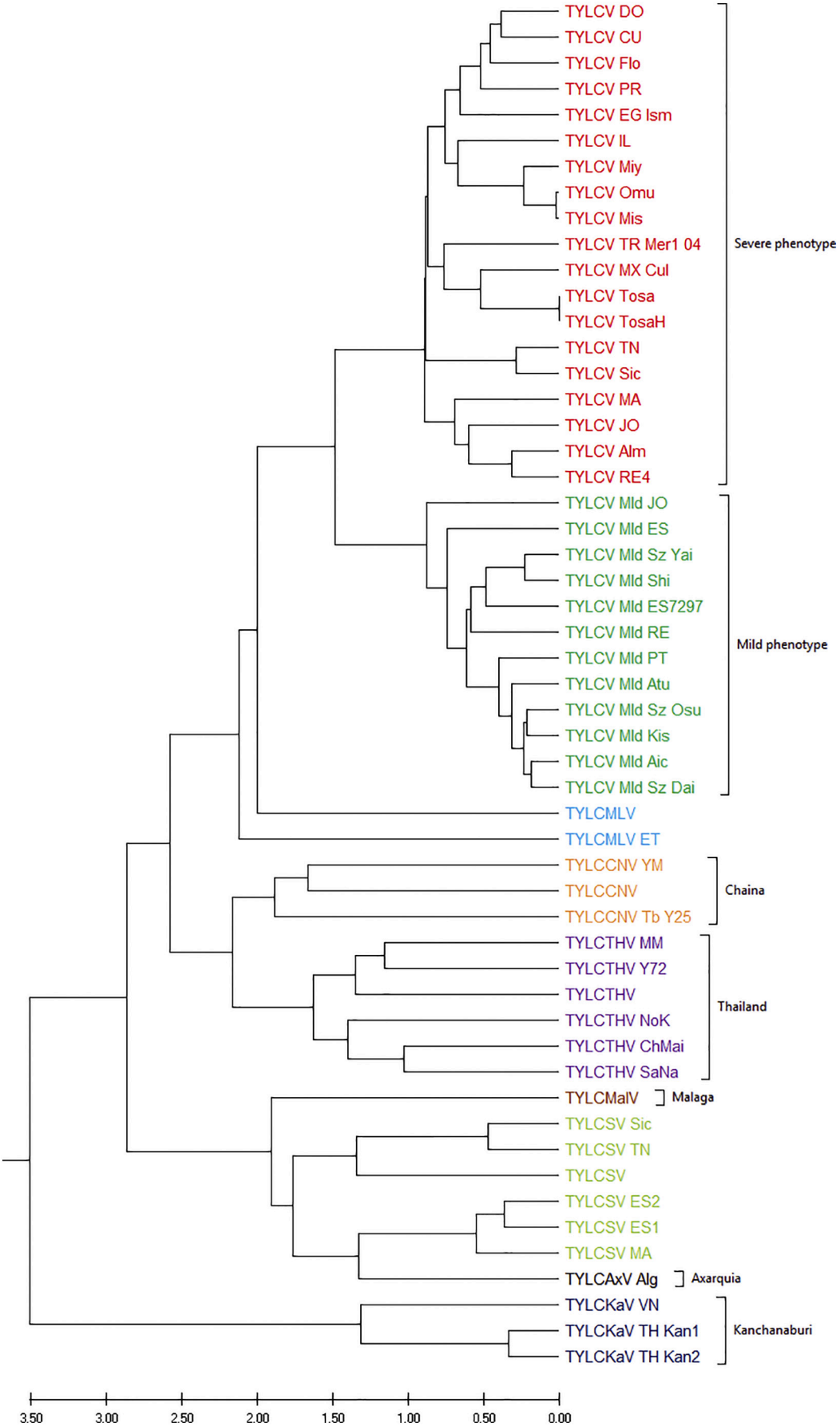

Fig. 7. Phylogenetic tree of 53 complete genome sequence of TYLCV based on our graph-theoretic approach.

JYO-HY003L(Japan), ARKELL (Canada), US1 (USA), US2 (USA), SWUS1(USA), HE-JA10 (Japan), JKN-SAP (Japan), JMY-HAW (Japan). First of all, we compare Fig. 6 with that of the Fuzzy integral similarity method [34] shown in supplementary Fig. S9. As per Fig. S9, US1 (US), which is originally from Genotype IV, comes under the group of Genotype III. Again SWJ570 (Japan), originally from Genotype IV, goes to Genotype III. This means that both of Genotype III and Genotype IV are not clustering properly. Therefore our method has a significant advantage, over the Fuzzy integral similarity method. We also compare Fig. 6 with Fig. S10 and Fig. S11 (shown in the supplementary file). These are obtained by a weighted measure method [42] and the ClustalW method respectively. It is observed that the HEV genotyping result based on weighted measure and ClustalW are well consistent with our proposed graph theoretical method. Finally, it may be stated that the result of our method is consistent with that of many other methods [43–48].

3.4. Analysis of 53 complete genome sequences of Tomato yellow leaf curl viruses (TYLCV) using Phylogenetic tree

We consider 53 complete genome sequences of Tomato yellow leaf curl virus (TYLCV) (shown in supplementary table S4). The minimum length of the sequence is 2731. 53 TYLCD-causing viral genomes are mainly of nine different types, TYLCV Severe phenotype (Red), Mild phenotype (Green), and the viruses from Axarquia (Black), Malaga (Brown), Mali (Sky Blue), Sardinia (Light Green), China (Orange), Kanchanaburi (Deep Blue), Thailand (Purple). For classification of TYLCV phylogenetic tree Fig. 7 is constructed using our graph-theoretic method.

The present result (Fig. 7) shows that all fifty-three viruses are distributed in nine distinct clusters. Our result is compared with the result obtained by the probabilistic method [23] shown in Fig. S12 (in the supplementary file). Fig. S12 shows that TYLCV Serve phenotype are scattered in three different clusters. Again Mild phenotype is also put in two different clusters. Lastly, six viruses of Sardinia are clustered with one Malaga and one Axarquia. This shows that the present method is better compared to the probabilistic method. Our present result is also on par with our previous result obtained by the k-mer method [29] shown in Fig. S13 (supplementary file).

3.5. Analysis of 59 bacterial genomes using phylogenetic tree

A large number of bacteria are found in the world, which is prokaryotic in nature. The length of the genome sequence of each bacterium is over 1 million (Mb). Therefore, a multiple sequence alignment method is not suitable to cluster them. So the problem is to develop methods to have their proper clustering and proper phylogeny. We consider the data set (SupplementaryTable S5) containing 59 bacterial genomes of 15 different families: Aeromonadaceae, Alcaligenaceae, Bacilleceae, Borreliaceae, Burkholderiaceae, Caulobacteraceae, Clostridiaceae, Desulfovibrionaceae, Enterobacteriaceae, Erwiniaceae, Lactobacillaceae, Mycoplasmataceae, Rhodobacteraceae, Staphylococcaceae, Yersiniaceae. Genomic sequence lengths in the data set vary from 3 to 10 MB. Now, we apply our graph-theoretic method to construct phylogeny on this dataset. It is shown in Fig. 8.

In Fig. 8, all the bacteria are classified into fifteen distinct clusters and further all the relative positions of the clusters are clearly shown therein. Now we compare our phylogeny with the phylogeny obtained by the multiple encoding vector method [40] shown in supplementary Fig. S14 and FFP method [30] shown in the supplementary Fig. S15. The three results are almost consistent with each other. However, in Fig. S15, the three families Lactobacillaceae, Clostridiaceae, and Staphylococcaceae from phylumBacilli are not clustered together as per closely related family. But In Fig. 8, all the bacteria are classified distinctly as per their family. Our result also agrees with that found in Fig. S14.

3.6. Analysis of 59 Ebola viruses using phylogenetic tree

For our next experiment, we consider the data set containing 59 complete genome sequences of five different types of viruses. These are Bundibugyo virus (BDBV), Reston virus (RESTV), Sudan virus (SUDV), Tai Forest virus (TAFV), and Ebola virus (formerly Zaire Ebolavirus, EBOV). These are given in supplementary file Table S6. EBOV viruses were found at different times in different places. Accordingly, they are designated differently. One is EBOV of Guinea in 2014; one type is called Gabon type, which occurred during 1994–1996, and also in 2002. One type is called DRC; it occurred in 2007, the same type of EBOV called Zaire (DRC) occurred during 1976–1977 and also in 1995. As a whole, we have six different types of Ebola virus in total. There are four other types of viruses, which are given by the names Bundibugyo virus (BDBV), Reston virus (RESTV), Sudan virus (SUDV), and Tai Forest virus (TAFV). Such 10 different classes of viruses are 59 in number and they are known by the general name of 59 Ebola viruses. The problem remains to show whether proper clustering in 10 separate clusters and the corresponding phylogeny can be obtained effectively. We construct a phylogenetic tree on this data set by our proposed graph-theoretic approach shown in Fig. 9.

As shown in Fig. 9, five different types of species are classified in distinct clusters. The Ebola viruses are totally separated from the other four (BDBV, RESTV, SUDV, and TAFV) viruses. Again EBOV stains are also separated into six different clusters. It is clearly shown in Fig. 9 that 20 different strains of EBOV 2014 outbreak in Guinea are grouped jointly. There are three different stresses of EBOV Zaire 1995 that are clustered together. The EBOV stresses in Zaire (DRC) [49] epidemic in 1976–1977 are collectively grouped together. The EBOV stresses in DRC endemic in 2007 are clustered jointly. Depending on the outbreak time, the EBOV which occurred in Gabon is separated into two dissimilar groups. The RESTV, SUDV, and BDBV are grouped separately amongst themselves; TAFV forms a group of singleton element. Overall phylogeny is also shown properly. Thus our method looks sound. Now we compare our phylogenetic tree (Fig. 9) with the phylogenetic tree, which was constructed by the FFP method [30] with k-mer length 7 as shown in Fig. S16 (supplementary file). As shown in Fig. S16, branch of SUDV and RESTV are not clustered together in a single branch. This is not consistent with the defined categorization of the Ebolavirus genus. For further comparison, we consider the phylogenetic tree, which was constructed using the Fuzzy integral similarity method [34], shown in Fig. S17 (supplementary file). Here, even six different types of EBOV are not clustered correctly. EBOV_2007_KC242788 is clustered with Zaire (DRC), 1976–1977 instead of DRC, 2007. Next, we compare the phylogenetic tree constructed by the ClustalW method shown in Fig. S18 (supplementary file). Here also branch of SUDV and RESTV are not clustered together in a single branch. Lastly, we compare our results with the phylogenetic tree (see supplementary Fig. S19) constructed using multiple encoding vector method [40]. As shown in Fig. S19, all six RESTV species are not clustered together. Thus so far as a phylogeny of 59 Ebolavirus is concerned, our method gives the best result.

3.7. Analysis of 38 influenza a virus using phylogenetic tree

Influenza disease is caused in birds and mammals [50] by the influenza A virus. Such viruses may be understood from their symbolic representation. We consider the dataset of 38 Influenza A viruses, which are described in Supplementary Table S7. A phylogenetic tree constructed by our method is shown in Fig. 10.

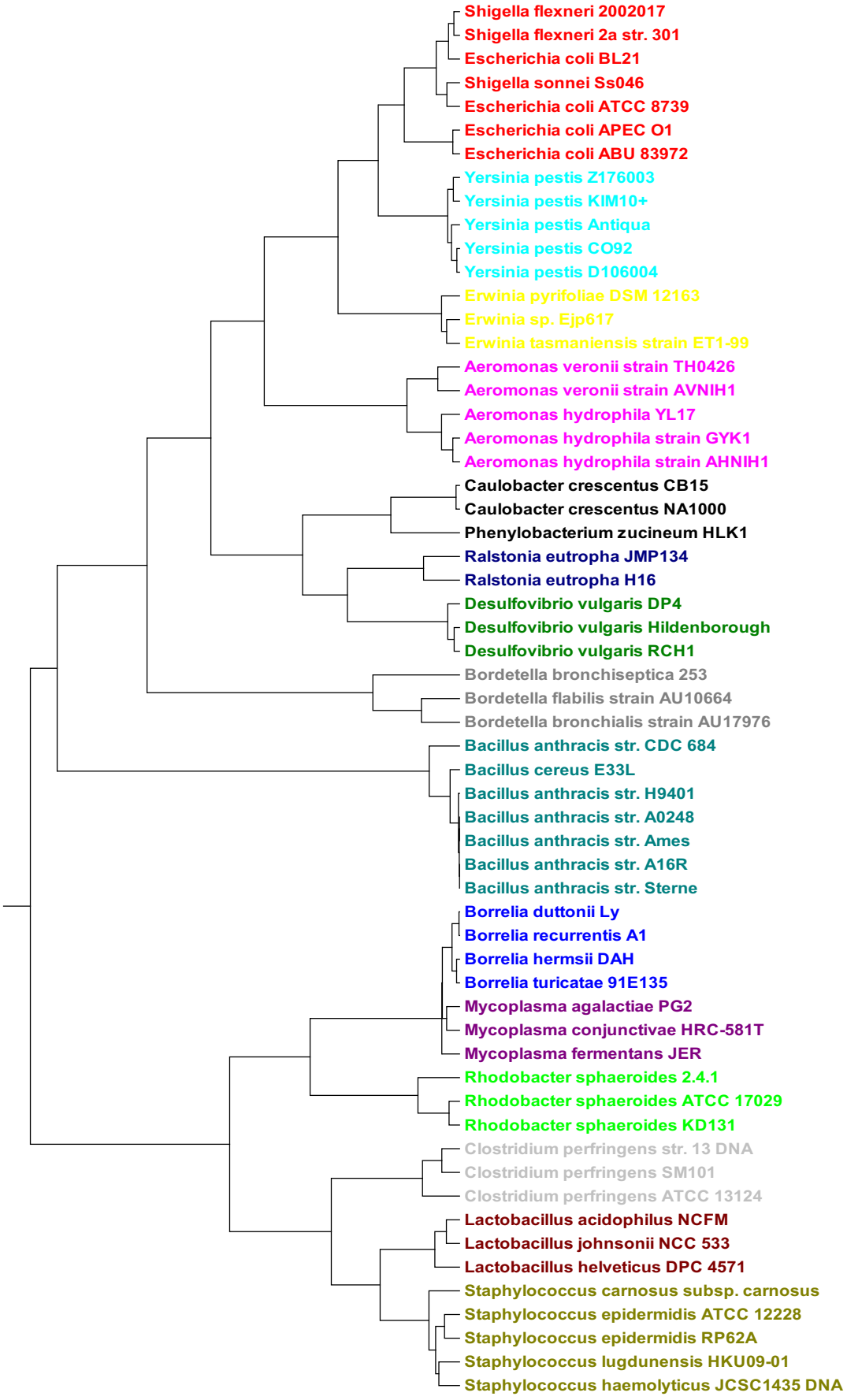

Fig. 8. Phylogenetic tree of 59 bacteria from 15 families under our graph-theoretic method.

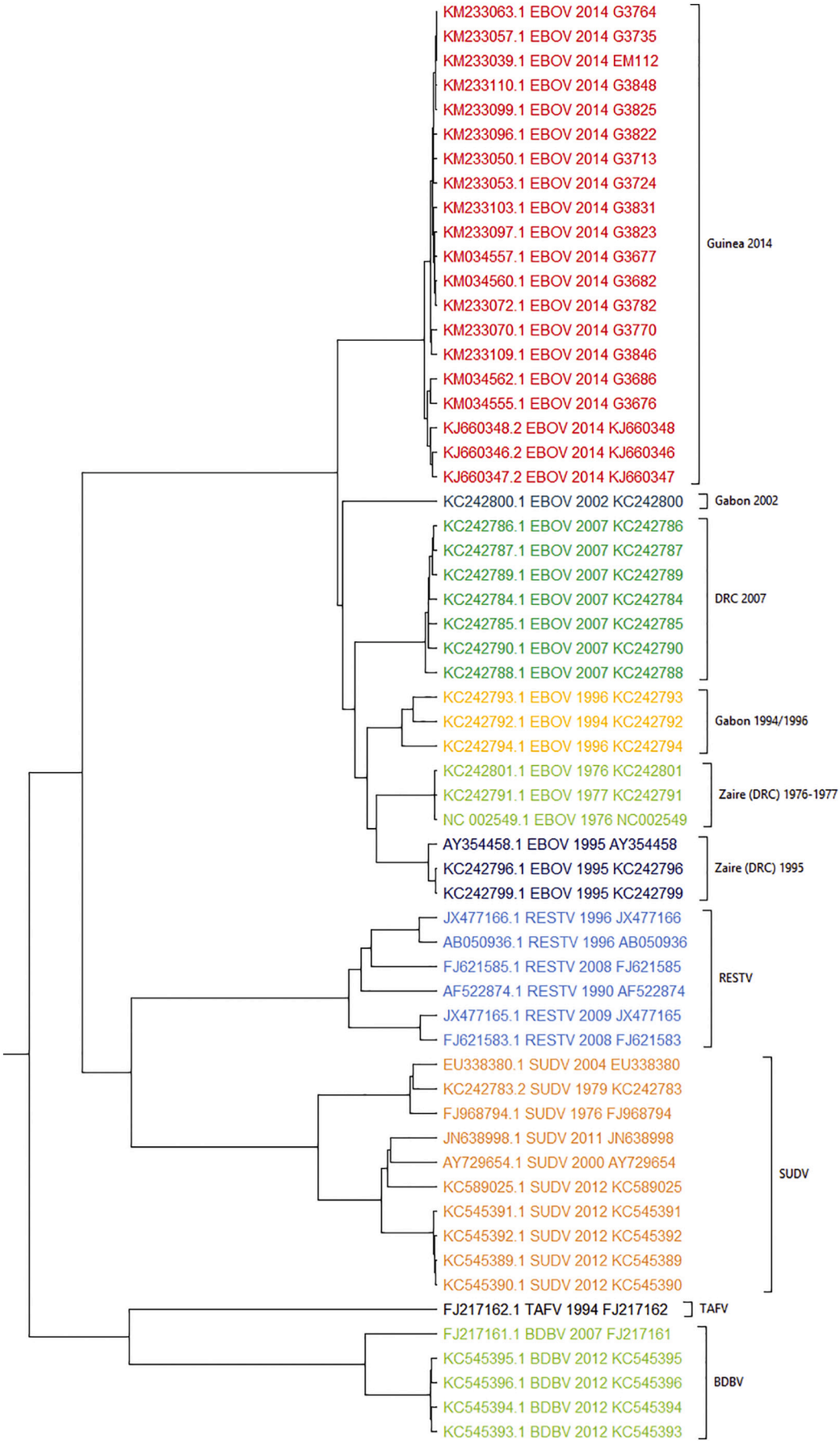

Fig. 9. Phylogenetic tree of 59 Ebolavirus genus based on our Graph-theoretic approach.

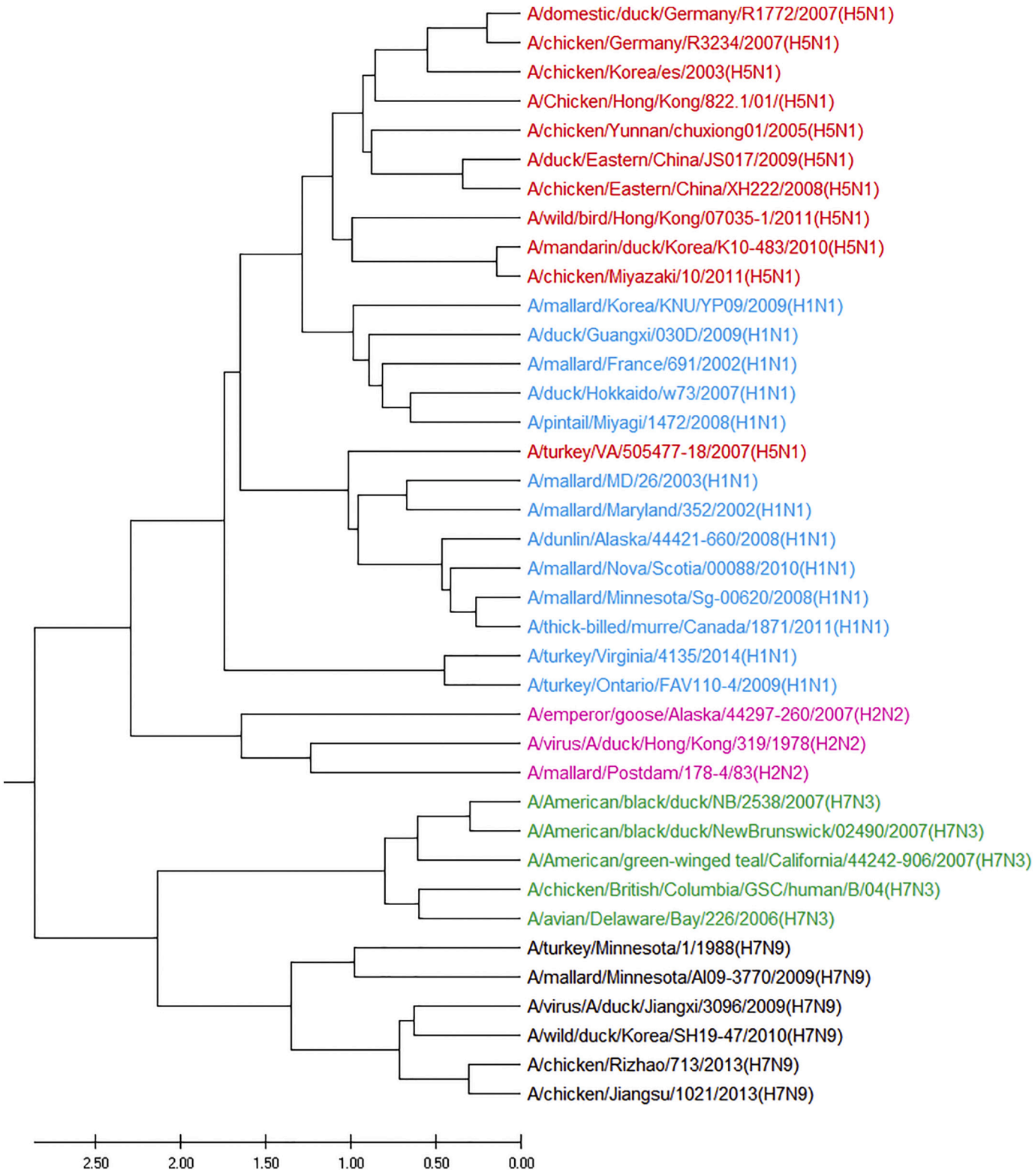

Fig. 10. Phylogenetic tree of 38 influenza-A viruses based on our graph-theoretic approach.

In our present method (Fig. 10), the 38 influenza-A viruses are correctly clustered into five groups but only one misclassification where A/Turkey/VA/505472–18/2007 (H5N1) is placed in the cluster of H1N1. Our result is consistent with the result obtained from the ClustalW method shown in Fig. S20 (in supplementary file). It also agrees with the result obtained by segment 6 of the k-mer method [51] shown in Fig. S21 (supplementary file) and another previous study [52]. Now Fig. S22 (supplementary file) obtained by FFP methods [30], where $k = 5$ shows misclassifications of several viruses between H5N1 and H5N5 clusters. Fig. S23 (supplementary file) obtained by multiple

vector encoding method [40] shows that all 38 viruses are clustered properly into five different clusters. Hence for the data set of 38 influenza-A viruses, multiple vector encoding method is by far the best and FFP method is the worst. Our method and the methods of ClustalW and segment 6 k-mer method are of the second-best choice.

4. Conclusion

In bioinformatics and evolutionary biology, genetic sequence comparison leads to an essential responsibility. To understand the relationship between different species, there are plenty of methods attempted by researchers. If the lengths of the genetic sequences are equal, alignment-based methods perform well. However, results are not reliable due to genetic retransformation and high metamorphosis rates. For a very large volume of genetic sequence data, these approaches are not at all suitable because of their computational difficulties. But alignment-free methods are more satisfactory since they decrease time complexity. Further, these approaches do not depend on the length of the sequences. Our proposed method is an alignment-free method based on a complete bipartite graph. Vertices of this graph are obtained under different combinations of mono-nucleotide and di-nucleotide of such sequences. This Complete Bipartite graph is a pioneering attempt in the representation of genetic sequence for analysis of their similarity/dissimilarity analysis. Our method is tested on the different categories of species including mammals, viruses (Ebola, Influenza, Hepatitis, TYLCV, Corona), and bacteria. On testing on those different datasets, our proposed graph theoretical method shows extremely good and accurate results over alignment-based as well as alignment-free methods.

Further, we calculate the time complexity of our proposed method, which is found to be O(N), whereas in alignment-based method, normally the time complexity found to be $O(N^2)$, N being the length of the sequence. So in comparison to time complexity, our proposed approach is much faster than the conventional alignment-based method. Also, no additional parameters like gap-counting, segment separation are required which are normally used in alignment-based methods to make the length of the sequences equal.

In conclusion, we state that we have found a new methodology based on graph theory to compare genetic data. Our proposed method is capable of presenting highly accurate evolutionary relationships amongst different types of species. The present method is very fast and appropriate for handling a large volume of the biological dataset.

Data availability

The datasets analyzed during the current study are available in the “Supplementary data".

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2020.08.023.

References

[1] A. Zielezinski, S. Vinga, J. Almeida, W.M. Karlowski, Alignment-free sequence comparison: benefits, applications, and tools, Genome Biol. 18 (1) (2017) 186.
[2] G. Bernard, C.X. Chan, Y.B. Chan, X.Y. Chua, Y. Cong, J.M. Hogan, ... M.A. Ragan, Alignment-free inference of hierarchical and reticulate phylogenomic relationships, Briefings in Bioinformatics 20 (2) (2019) 426–435.
[3] W. Just, Computational complexity of multiple sequence alignment with SP-score, J. Comput. Biol. 8 (6) (2001) 615–623.
[4] S. Vinga, J. Almeida, Alignment-free sequence comparison—a review, Bioinformatics 19 (4) (2003) 513–523.
[5] S. Das, S. Ghosh, J. Pal, D.K. Bhattacharya, Use of fuzzy set theory in DNA sequence comparison and amino acid classification, Emerging Research on Applied Fuzzy Sets and Intuitionistic Fuzzy Matrices, IGI Global, 2017, pp. 235–253.
[6] T. Hoang, C. Yin, S.S.T. Yau, Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison, Genomics 108 (3–4) (2016) 134–142.
[7] S. Das, A. Das, B. Mondal, N. Dey, D.K. Bhattacharya, D.N. Tibarewala, Genome sequence comparison under a new form of tri-nucleotide representation based on bio-chemical properties of nucleotides, Gene 730 (2020) 144257.
[8] Y. Wu, A.W.C. Liew, H. Yan, M. Yang, DB-curve: a novel 2D method of DNA sequence visualization and representation, Chem. Phys. Lett. 367 (1–2) (2003) 170–176.
[9] B. Liao, X. Xiang, W. Zhu, Coronavirus phylogeny based on 2D graphical representation of DNA sequence, J. Comput. Chem. 27 (11) (2006) 1196–1202.
[10] S. Das, S. Palit, A.R. Mahalanabish, N.R. Choudhury, A new way to find similarity/dissimilarity of DNA sequences on the basis of dinucleotides representation, Computational Advancement in Communication Circuits and Systems, Springer, New Delhi, 2015, pp. 151–160.
[11] M. Randić, M. Vracko, A. Nandy, S.C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization, J. Chem. Inf. Comput. Sci. 40 (5) (2000) 1235–1244.
[12] B. Liao, T.M. Wang, 3-D graphical representation of DNA sequences and their numerical characterization, J. Mol. Struct. THEOCHEM 681 (1–3) (2004) 209–212.
[13] S. Das, N.R. Choudhury, D.N. Tibarewala, D.K. Bhattacharya, Application of Chaos game in tri-nucleotide representation for the comparison of coding sequences of β-Globin Gene, Industry Interactive Innovations in Science, Engineering and Technology, Springer, Singapore, 2018, pp. 561–567.
[14] X. Zhang, J. Luo, L. Yang, New invariant of DNA sequence based on 3DD-curves and its application on phylogeny, J. Comput. Chem. 28 (14) (2007) 2342–2346.
[15] X.Q. Qi, J. Wen, Z.H. Qi, New 3D graphical representation of DNA sequence based on dual nucleotides, J. Theor. Biol. 249 (4) (2007) 681–690.
[16] P. Wąż, D. Bielińska-Wąż, 3D-dynamic representation of DNA sequences, J. Mol. Model. 20 (3) (2014) 2141.
[17] M. Randić, A.T. Balaban, On a four-dimensional representation of DNA primary sequences, J. Chem. Inf. Comput. Sci. 43 (2) (2003) 532–539.
[18] R. Chi, K. Ding, Novel 4D numerical representation of DNA sequences, Chem. Phys. Lett. 407 (1–3) (2005) 63–67.
[19] C. Tan, S. Li, P. Zhu, 4D graphical representation research of DNA sequences, Int. J. Biomath. 8 (01) (2015) 1550004.
[20] Z. Mo, W. Zhu, Y. Sun, Q. Xiang, M. Zheng, M. Chen, Z. Li, One novel representation of DNA sequence based on the global and local position information, Sci. Rep. 8 (1) (2018) 1–7.
[21] B. Liao, Q. Xiang, L. Cai, Z. Cao, A new graphical coding of DNA sequence and its similarity calculation, Phys. A Stat. Mech. Appl. 392 (19) (2013) 4663–4667.
[22] M. Randić, On characterization of DNA primary sequences by a condensed matrix, Chem. Phys. Lett. 317 (1–2) (2000) 29–34.
[23] C. Yu, M. Deng, S.S.T. Yau, DNA sequence comparison by a novel probabilistic method, Inf. Sci. 181 (8) (2011) 1484–1492.
[24] I. Schwende, T.D. Pham, Pattern recognition and probabilistic measures in alignment-free sequence analysis, Brief. Bioinform. 15 (3) (2014) 354–368.
[25] Y. Li, T. Song, J. Yang, Y. Zhang, J. Yang, An alignment-free algorithm in comparing the similarity of protein sequences based on pseudo-markov transition probabilities among amino acids, PLoS One 11 (12) (2016) e0167430.
[26] X. Yang, T. Wang, A novel statistical measure for sequence comparison on the basis of k-word counts, J. Theor. Biol. 318 (2013) 91–100.
[27] C. Li, Y. Yang, M. Jia, Y. Zhang, X. Yu, C. Wang, Phylogenetic analysis of DNA sequences based on k-word and rough set theory, Phys. A Stat. Mech. Appl. 398 (2014) 162–171.
[28] J. Wen, R.H. Chan, S.C. Yau, R.L. He, S.S. Yau, K-mer natural vector and its application to the phylogenetic analysis of genetic sequences, Gene 546 (1) (2014) 25–34.
[29] S. Das, T. Deb, N. Dey, A.S. Ashour, D.K. Bhattacharya, D.N. Tibarewala, Optimal choice of k-mer in composition vector method for genome sequence comparison, Genomics 110 (5) (2018) 263–273.
[30] G.E. Sims, S.R. Jun, G.A. Wu, S.H. Kim, Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions, Proc. Natl. Acad. Sci. 106 (8) (2009) 2677–2682.
[31] X. Wu, X.F. Wan, G. Wu, D. Xu, G. Lin, Phylogenetic analysis using complete signature information of whole genomes and clustered neighbour-joining method, Int. J. Bioinforma. Res. Appl. 2 (3) (2006) 219–248.
[32] L. Gao, J. Qi, Whole genome molecular phylogeny of large dsDNA viruses using composition vector method, BMC Evol. Biol. 7 (1) (2007) 41.
[33] G. Lu, S. Zhang, X. Fang, An improved string composition method for sequence comparison, BMC Bioinforma. 9 (S6) (2008) S15.
[34] A.K. Saw, B.C. Tripathy, S. Nandi, Alignment-free similarity analysis for protein sequences based on fuzzy integral, Sci. Rep. 9 (1) (2019) 1–13.
[35] X. Qi, Q. Wu, Y. Zhang, E. Fuller, C.Q. Zhang, A novel model for DNA sequence similarity analysis based on graph theory, Evol. Bioinform. 7 (2011) EBO-S7364.
[36] R. Mathur, N. Adlakha, A graph theoretic model for prediction of reticulation events and phylogenetic networks for DNA sequences, Egypt. Basic Appl. Sci. 3 (3) (2016) 263–271.
[37] J.A. Bondy, U.S.R. Murty, Graph Theory with Applications, vol. 290, Macmillan, London, 1976.
[38] S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, Mol. Biol. Evol. 33 (7) (2016) 1870–1874.
[39] W.M. Brown, E.M. Prager, A. Wang, A.C. Wilson, Mitochondrial DNA sequences of

primates: tempo and mode of evolution, J. Mol. Evol. 18 (4) (1982) 225–239.
[40] Y. Li, L. He, R.L. He, S.S.T. Yau, A novel fast vector method for genetic sequence comparison, Sci. Rep. 7 (1) (2017) 1–11.
[41] C. Yu, Q. Liang, C. Yin, R.L. He, S.S.T. Yau, A novel construction of genome space with biological geometry, DNA Res. 17 (3) (2010) 155–168.
[42] L. Liu, C. Li, F. Bai, Q. Zhao, Y. Wang, An optimization approach and its application to compare DNA sequences, J. Mol. Struct. 1082 (2015) 49–55.
[43] Z. Liu, J. Meng, X. Sun, A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping, Biochem. Biophys. Res. Commun. 368 (2) (2008) 223–230.
[44] S. Ding, Q. Dai, H. Liu, T. Wang, A simple feature representation vector for phylogenetic analysis of DNA sequences, J. Theor. Biol. 265 (4) (2010) 618–623.
[45] Y. Huang, L. Yang, T. Wang, Phylogenetic analysis of DNA sequences based on the generalized pseudo-amino acid composition, J. Theor. Biol. 269 (1) (2011) 217–223.
[46] X. Yang, T. Wang, A novel statistical measure for sequence comparison on the basis of k-word counts, J. Theor. Biol. 318 (2013) 91–100.
[47] W. Hou, Q. Pan, M. He, A novel representation of DNA sequence based on CMI coding, Phys A Stat. Mech. Appl. 409 (2014) 87–96.
[48] L. Liu, C. Li, F. Bai, Q. Zhao, Y. Wang, An optimization approach and its application to compare DNA sequences, J. Mol. Struct. 1082 (2015) 49–55.
[49] E.C. Holmes, G. Dudas, A. Rambaut, K.G. Andersen, The evolution of Ebola virus: insights from the 2013–2016 epidemic, Nature 538 (7624) (2016) 193–200.
[50] D. Vijaykrishna, L.L. Poon, H.C. Zhu, et al., Reassortment of pandemic H1N1/2009 influenza A virus in swine, Science 328 (5985) (2010) 1529, https://doi.org/10.1126/science.1189132.
[51] T. Hoang, C. Yin, H. Zheng, C. Yu, R.L. He, S.S.T. Yau, A new method to cluster DNA sequences using Fourier power spectrum, J. Theor. Biol. 372 (2015) 135–145.
[52] C. Yin, S.S.T. Yau, An improved model for whole genome phylogenetic analysis by Fourier transform, J. Theor. Biol. 382 (2015) 99–110.