SCIENTIFIC REPORTS

OPEN

Alignment-free method for DNA sequence clustering using Fuzzy integral similarity

Received: 31 July 2018
Accepted: 28 January 2019
Published online: 06 March 2019

Ajay Kumar Saw[1], Garima Raj[2], Manashi Das[2], Narayan Chandra Talukdar[2], Binod Chandra Tripathy[3] & Soumyadeep Nandi[2]

A larger amount of sequence data in private and public databases produced by next-generation sequencing put new challenges due to limitation associated with the alignment-based method for sequence comparison. So, there is a high need for faster sequence analysis algorithms. In this study, we developed an alignment-free algorithm for faster sequence analysis. The novelty of our approach is the inclusion of fuzzy integral with Markov chain for sequence analysis in the alignment-free model. The method estimate the parameters of a Markov chain by considering the frequencies of occurrence of all possible nucleotide pairs from each DNA sequence. These estimated Markov chain parameters were used to calculate similarity among all pairwise combinations of DNA sequences based on a fuzzy integral algorithm. This matrix is used as an input for the neighbor program in the PHYLIP package for phylogenetic tree construction. Our method was tested on eight benchmark datasets and on in-house generated datasets (18 s rDNA sequences from 11 arbuscular mycorrhizal fungi (AMF) and 16 s rDNA sequences of 40 bacterial isolates from plant interior). The results indicate that the fuzzy integral algorithm is an efficient and feasible alignment-free method for sequence analysis on the genomic scale.

Phylogenetic tree analysis and comparative studies of taxa are essential parts of modern molecular biology. Phylogenetic reconstruction and comparative sequence analysis traditionally depend on multiple or pairwise sequence alignments. However, various limitations are encountered when analyzing large datasets using an alignment based approach. Whole genome alignment of higher eukaryotes can exceed computational resources. Moreover, factors such as the combinatorics of genomic rearrangements and duplications make the alignment of entire genomes impossible. Therefore, the alignable homologous segments of the genomes under study have to be identified in the initial steps. Recently, large amounts of sequence data produced by next-generation sequencing techniques have become available in private and public databases, which has created new challenges due to the limitations associated with alignment based approaches. This plethora of sequence information increases the computation and time requirements for genome comparisons in computational biology. Therefore, there is a high need for faster sequence analysis algorithms. For this, various methods have been proposed to overcome the limitations of alignment based approach[1–3], and is termed as alignment-free methods. The alignment-free methods are not only used in phylogenetic studies[4,5], but also for metagenomics[6–11], analysis of regulatory elements[12–14], protein classification[15,16], sequence assembly[17], isoform quantification from transcriptome data[18], and to identify biomarkers in diagnostic tests[19]. The alignment-free methods fall into two broad categories: methods based on k-mer or word frequency, and methods based on match length[20]. Methods based on k-mer or word frequency are quite popular and studied extensively. The k-mer based methods were developed to compare DNA sequences, in which it counts the frequencies of substrings with k letters occurring in respective sequences[21]. In recent past, a lot of k-mer based methods have been proposed and implemented in sequence analysis and phylogeny, such as, feature frequency profile (FFP)[22], return time distribution (RTD)[23], frequency chaos game representation (FCGR)[24], an improved complete composition vector method (ICCV)[25], composition vector (CV)[26] and complete composition vector (CCV)[27]. For sequence comparison, ICCV method is more efficient and robust compared to CV and CCV methods. The other category of the alignment-free method is based on match lengths,

[1]Institute of Advanced Study in Science and Technology, Mathematical Sciences Division, Guwahati, 781035, India. [2]Institute of Advanced Study in Science and Technology, Life Science Division, Guwahati, 781035, India. [3]Tripura University, Department of Mathematics, Agartala, 799022, India. Correspondence and requests for materials should be addressed to S.N. (email: soumyadeep.nandi@gmail.com)

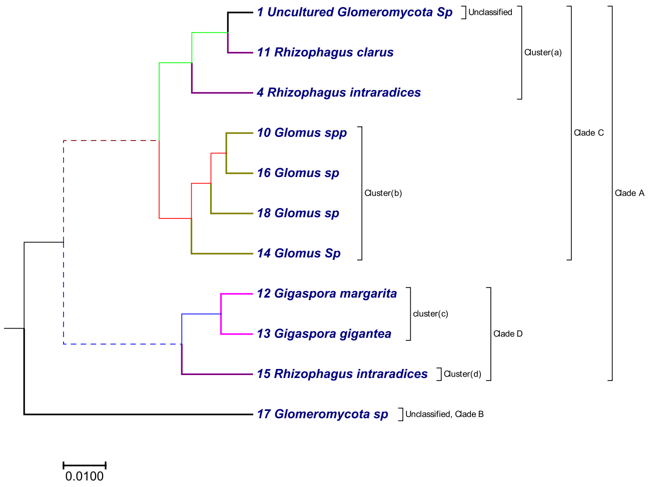


Figure 1. The phylogenetic tree of the 11 AMF sequences constructed using our method.

where it employs the similarity of substrings between two sequences[28–31]. Examples of match length methods are, k-mismatch average common substring[32], average common substring[28], K_r – method[28], etc. These methods are commonly used for string processing in computer science. In this study, we propose to use fuzzy integral[33] to analyze DNA sequences based on a Markov chain[34], which can be categorised as k-mer or word frequency method. The fuzzy integral similarity method[35,36] assigns a similarity score between two DNA sequences based on the estimated parameters of a Markov chain. A DNA sequence consists of four characters (A, T, G and C). By taking the state space as **S** = {A, T, G, C}, we used the *k*-th step transition probability matrix, a fuzzy measure[37] and fuzzy integral to describe the DNA sequences. We used the fuzzy integral similarity to obtain a distance matrix, which was used in the neighbor program in the PHYLIP package[38] to construct a phylogenetic tree. The similar fuzzy integral similarity approach was taken by[36]. However in[36], the method of feature vector extraction from the DNA sequences is different from our method. In both our method and[36], the extracted features are used as an input for the fuzzy integral similarity analysis. The proposed method is tested on 18S rDNA sequences from 11 Arbuscular mycorrhizal fungi isolates and 16S rDNA sequences from 40 bacterial isolates, and also tested on the following benchmark datasets, 41 mammalian mitochondrial genomes, 59 ebolavirus complete genomes, 30 coronavirus whole genomes, 30 bacterial whole genomes, 48 Hepatitis E virus (HEV) whole genomes, 24 Eutherian mammals sequences, 58 genome datasets from different species and 29 Escherichia/Shigella complete genomes. The method was also tested on large mammalian dataset. In addition, we used receiver operating characteristic (ROC)[39–41] curve for measuring the performance of our method to compare the other alignment-free methods from Alfree repository[2]. The consistency can also be seen from the statistical analysis such as AUC (area under the ROC) values, calculated from ROC curves provided in Supplementary Material.

Materials and Methods

Construction of a Markov chain for DNA sequence.

Let $P = [p_{ij}]$ denote the transition probability matrix of a discrete-time Markov chain[34]. Each state transition probability p_{ij} is defined as follows:

$$p_{ij} = p(X_{n+1} = s_j | X_n = s_i), \quad 1 \leq i, j \leq S, \tag{1}$$

where X_n indicates the actual state at time $n(n = 1, 2, 3 \ldots)$ and s_i is the i_{th} state of S distinct states. In the context of a DNA sequence, the number of states is $S = 4$ which corresponds to the four nucleotide symbol set $\mathbf{S} = \{A = s_1, T = s_2, G = s_3, C = s_4\}$. The state transition probabilities are subject to

$$p_{ij} \geq 0 \ \ \forall\, i, j \ \text{ and } \sum_{j=1}^{S} p_{ij} = 1 \ \ \forall\, i.$$

Since the transition probabilities are unknown initially, they must be estimated based on the observed sequence. Here, we estimate the parameters of the Markov chain by taking the frequencies of occurrence of all possible nucleotide pairs for each sequence[42]. If the total number of each adjacent nucleotide pair (s_i, s_j) in the sequence is denoted by $N_{s_i s_j}$, then the 1^{st}-step transition probability from state s_i to state s_j is estimated as

$$p_{ij} = \frac{N_{s_i s_j}}{N_{s_i A} + N_{s_i T} + N_{s_i G} + N_{s_i C}}, \tag{2}$$

where $N_{s_i s_j}$ represents the total number of each adjacent pair starting from nucleotide s_i and ending with nucleotide s_j.

Presented above is the 1-step Markov chain. The k-step Markov chain can be calculated through the 1-step Markov chain, which is known as the Chapman-Kolmogorov process. Let $P^k = [p_{ij}^k]$ denote the transition probability matrix of a discrete-time Markov chain in state j after k steps from state i. Each state transition probability p_{ij}^{k} is defined as follows:

$$p_{ij}^{k}=p^{k}(X_{n+k}=s_j|X_n=s_i),\quad 1\le i,j\le S, \tag{3}$$

The state transition probabilities are subject to

$$p_{ij}^{k}\ge 0\ \ \forall\, i,j\ \text{ and }\sum_{j=1}^{S}p_{ij}^{k}=1\ \ \forall\, i.$$

For any three events, A, B and C, the following identity is known: $p[A\cap B|C]=p[A|B\cap C]p[B|C]$. By interpreting A as $X_{n+k}=s_j$, B as $X_{n+t}=s_r$ and C as $X_n=s_i$, we have

$$\begin{aligned}p_{ij}^{k}&=p[X_{n+k}=s_j|X_n=s_i]\\&=\sum_{s_r\in\mathbf{S}}p[X_{n+k}=s_j,X_{n+t}=s_r|X_n=s_i]\\&=\sum_{s_r\in\mathbf{S}}p[X_{n+k}=s_j|X_{n+t}=s_r,X_n=s_i]\times p[X_{n+t}=s_r|X_n=s_i]\\&=\sum_{s_r\in\mathbf{S}}p[X_{n+k}=s_j|X_{n+t}=s_r]\times p[X_{n+t}=s_r|X_n=s_i]\\&=\sum_{s_r\in\mathbf{S}}p_{rj}^{k-t}p_{ir}^{t},\end{aligned} \tag{4}$$

which is known as the Chapman-Kolmogorov equation.

Hence, the matrix with the elements p_{ij}^{k} is $[p_{ij}^{k}]=P^{k}$.

The selection of step k plays an important role in capturing rich evolutionary information from the DNA sequence. In the context of a DNA sequence, the k^{th}-step transition probability can be written as:

$$P^{k}=\begin{bmatrix}p_{11}^{k}&p_{12}^{k}&p_{13}^{k}&p_{14}^{k}\\p_{21}^{k}&p_{22}^{k}&p_{23}^{k}&p_{24}^{k}\\p_{31}^{k}&p_{32}^{k}&p_{33}^{k}&p_{34}^{k}\\p_{41}^{k}&p_{42}^{k}&p_{43}^{k}&p_{44}^{k}\end{bmatrix}=\begin{bmatrix}p_{AA}^{k}&p_{AT}^{k}&p_{AG}^{k}&p_{AC}^{k}\\p_{TA}^{k}&p_{TT}^{k}&p_{TG}^{k}&p_{TC}^{k}\\p_{GA}^{k}&p_{GT}^{k}&p_{GG}^{k}&p_{GC}^{k}\\p_{CA}^{k}&p_{CT}^{k}&p_{CG}^{k}&p_{CC}^{k}\end{bmatrix} \tag{5}$$

Which is subject to $p_{ij}^{k}\ge 0\ \ \forall\, i,\ j\in\{1,2,3,4\}$ and $\sum_{j=1}^{4}p_{ij}^{k}=1\ \ \forall\, i$. The p_{ij}^{k} can be calculated using the above Eqs (2 and 4).

Fuzzy measure and fuzzy integral for the k^{th}-step nucleotide sequence.

Fuzzy set theory[43] is particularly suitable for modelling imprecise data, whereas fuzzy integral is highly appropriate for representing the interaction among different information sources. The concept of fuzzy integral with respect to a fuzzy measure has been proposed by Sugeno in 1974[44]. In this section, we propose the use of the fuzzy integral incorporating with the transition probability matrix, where the elements of transition probability matrix are taken as fuzzy membership degree.

Let $F=\{(s_is_j)^{k}=y_{ij}|i,j\in\{1,2,3,4\}\}$ be the finite set of k^{th}-step nucleotides starting from nucleotide s_i and ending with nucleotide s_j estimated from the observed sequence.

Let X, $Y\subseteq F$ and $R(F)$ be the power set of F. A fuzzy measure μ is a real valued function:

μ: $R(F)\to[0,1]$ satisfies the given condition,

(a) $\mu(\phi)=0$ and $\mu(F)=1$.
(b) $\mu(X)\le\mu(Y)$ if $X\subseteq Y$.

For a fuzzy measure μ let $\mu(y_{ij})=\mu^{ij}\ \ \forall\, y_{ij}\in F$. The mapping $y_{ij}\to\mu^{ij}$ is termed a fuzzy density function. The fuzzy density function can be interpreted as the importance of element y_{ij} in determining the set F. By definition of the fuzzy measure μ, the measure of the union of two disjointed subsets cannot be directly computed from their disjointed component measures. In other words, the fuzzy measure value of a given subset is not simply the summation of the measures of its elements. Therefore, to define a fuzzy measure, we need to know the fuzzy densities of each element of the measured set and the measure of each combination. This measure can be provided by an expert or extracted from the problem definition. However, when dealing with a set of numerous elements, this job may become noisy, tedious or even unfeasible. A possible solution to this problem is to use a λ – fuzzy measure. A λ – fuzzy measure[45] fulfills the criteria of a fuzzy measure, and has an additional property: for all $X\cap Y=\phi,\ X,Y\subseteq\{y_{i1},y_{i2},y_{i3},y_{i4}\}$ for fixed $i\in\{1,2,3,4\}$ and

$$\mu(X\bigcup Y)=\mu(X)+\mu(Y)+\lambda_i\mu(X)\mu(Y),\ \text{ for each }\lambda_i>-1. \tag{6}$$

Furthermore, λ_i can be calculated by solving:

$$\lambda_i + 1 = \prod_{j=1}^{4} (1 + \lambda_i \mu^{ij}) \text{ for fixed } i. \tag{7}$$

For solving Eqs (6 and 7), we only need to assemble information regarding the individual fuzzy densities of the elements μ^{ij} $(i, j = 1, 2, 3, 4)$.

Let $F = \{(s_i s_j)^k = y_{ij} | i, \ j \in \{1, 2, 3, 4\}\}$ be a finite set of information sources. Let h: $F \rightarrow [0, 1]$ represent a function that maps each element of F to its observed evidence. Suppose that $h(y_{i1}) \geq h(y_{i2}) \geq h(y_{i3}) \geq h(y_{i4})$ for each fixed $i \in \{1, 2, 3, 4\}$ If the decreasing order criterion is not fulfilled, then F should be reordered so that the decreasing order relationship holds, and further investigation will be based on the modified relationship. Let μ: $R(F) \rightarrow [0, 1]$ be a fuzzy measure. Then, the fuzzy integral of h with respect to the fuzzy measure μ is

$$I = max[max_{i=1}^{4}[min_{j=1}^{4}[h(y_{ij}), \mu(A_{ij})]]], \tag{8}$$

$$\text{where } A_{ij} = \{y_{i1}, y_{i2}, \ldots, y_{ij}\} \text{ for each fixed } i. \tag{9}$$

The fuzzy integral considers the significance provided by every element of a given set, and the importance of each subset of elements (i.e., the fuzzy measure) plays an important role in its decision-making process. The combination of the extracted information and the importance of the provided source convert the fuzzy integral to an appropriate form for information fusion. This theory has the potential to address uncertainties associated with issues related to data extraction and their processing procedures. Therefore, the theory has been widely implemented in feature extraction and classification[45,46].

Fuzzy integral similarity and distance matrix for sequence comparison. The fuzzy integral similarity is based on the distance of the k^{th}-step nucleotide pair frequency with respect to the conservation level of the position between two sequences. In our case, the k^{th}-step nucleotide pair frequency at all sixteen positions in the transition probability matrix is taken as the fuzzy membership degree.

Let P_1^k and P_2^k be two k^{th}-step transition probability matrices. The fuzzy integral function find the similarity level of the nucleotide pairs between k^{th}-step transition probability matrices. We constructed a fuzzy integral function h, which is given as:

$$h^{y_{ij}} = 1 - |(P_1^k)^{y_{ij}} - (P_2^k)^{y_{ij}}|, \tag{10}$$

where $y_{ij} \in \{(AA)^k, (AT)^k, (AG)^k, (AC)^k, (TA)^k, (TT)^k, (TG)^k, (TC)^k, (GA)^k, (GT)^k, (GG)^k, (GC)^k, (CA)^k, (CT)^k, (CG)^k, (CC)^k\}$.

Additionally, the fuzzy measure function find the maximum level of conservation of the nucleotide pairs between k^{th}-step transition probability matrices P_1^k and P_2^k, which favours the importance of better conserved positions.

Taking advantage of the properties explained above, we can construct a λ – fuzzy measure μ using the fuzzy density of each element μ^{ij}.

In this case,

$$\mu^{ij} = \mu^{y_{ij}} = max\{(P_1^k)^{y_{ij}}, (P_2^k)^{y_{ij}}\}, \tag{11}$$

where $y_{ij} \in \{(AA)^k, (AT)^k, (AG)^k, (AC)^k, (TA)^k, (TT)^k, (TG)^k, (TC)^k, (GA)^k, (GT)^k, (GG)^k, (GC)^k, (CA)^k, (CT)^k, (CG)^k, (CC)^k\}$. At this stage, we should apply Eq. (7) to find λ and apply the value of λ in Eq. (6) to finally obtain the fuzzy measure μ. The result generated by Eq. (6) satisfies the given criteria (a) and (b) of the fuzzy measure. After generating h and μ, we obtained the fuzzy integral similarity by applying Eq. (8). In fuzzy integral similarity, greater importance is given to the higher degree of membership which is calculated via the fuzzy integral with respect to the fuzzy measure. It is based on fuzzy technology and is intended to deal with the intrinsic uncertainty involved in sequence comparison tasks. Fuzzy integral similarity does not require any additional parameters, which makes it fully automated and robust.

The fuzzy integral similarity measure provides the similarity score between the two k^{th}-step transition probability matrices. Next, we will define a distance measure between two k^{th}-step transition probability matrices P_1^k and P_2^k, which is given as follows:

$$D(P_1^k, P_1^k) = 1 - I(P_1^k, P_1^k). \tag{12}$$

Similarly using Eq. (12), we can calculate the distance measure for all pairwise combinations taken from an n number of DNA sequences. Finally, a symmetric distance matrix is generated. This matrix is used as an input for the neighbor program in the PHYLIP package[38] for phylogenetic tree construction.

Algorithm. This section describes the algorithmic aspect of the proposed method. The entire algorithm contains three stages.

Stage 1: Calculation of the transition probability matrix through Markov chain:

Algorithm 1. k^{th}-step transition probability matrix.

```
read all DNA sequences D from input file
bases ← $\{s_1 = A, s_2 = T, s_3 = G, s_4 = C\}$
while (D!=NULL) do
    l ← string length of DNA
    set DNA array D[l]
    repeat all possible base pairs do
        g ,q ∈ bases
        set n[g][q] ← 0
        for($i = 0$ to $i = l - 2$ $step$) do
            $j \leftarrow i + 1$
            if(D[i]==g and D[j]==q) do
                n[g][q] ← n[g][q]+1
            end if
        end for
    end repeat
    for $(1 <= i, j <= 4)$ do
        store n[$s_i$][$s_j$]
        $p[i][j] \leftarrow n[s_i][s_j] / \sum_{j=1}^{4} n[s_i][s_j]$
    end for
    $P \leftarrow [p_{ij}]_{1<=i,j<=4}$
    $P^1 \leftarrow P$   for $1^{st}$-step
    $P^2 \leftarrow P * P$   for $2^{nd}$-step
    return $P^k \leftarrow P * P..(k - times)..P$ for $k^{th}$-step
end while
```

Stage 2: Calculation of fuzzy integral similarity between two $k^{th}-$ step transition probability matrices:

Algorithm 2. FISim (P_1^k, P_2^k).

```
input $P_1^k, P_2^k$
for $i \in \{1, 2, 3, 4\}$ do
    for $j \in \{1, 2, 3, 4\}$ do
        $h(y_{ij}) \leftarrow\ 1 - |y_{ij}^{P_1^k} - y_{ij}^{P_2^k}|$
        $\mu(y_{ij}) \leftarrow\ max(y_{ij}^{P_1^k}, y_{ij}^{P_2^k})$
    end for
    $\{h(y_{ij})\}_{j=1}^{4} \leftarrow$ sort $\{h(y_{ij})\}_{j=1}^{4}$ in decreasing order
    $\{\mu(y_{ij})\}_{j=1}^{4} \leftarrow \{\mu(y_{ij})\}_{j=1}^{4}$ positional arrangement based on $\{h(y_{ij})\}_{j=1}^{4}$
    $\lambda_i \leftarrow$ solve equation (7) using the Newton Rapson method
    for $(j \in \{1, 2, 3, 4\})$ do
        $A_{ij} \leftarrow \{y_{i1}, y_{i2}, ..., y_{ij}\}$
        $\mu(A_{ij}) \leftarrow$ calculate equation(6)
    end for
    $J_i \leftarrow max_i[min_{j=1}^{4}[h(y_{ij}), \mu(A_{ij})]]$
end for
FISim$(P_1^k, P_2^k) \leftarrow max_{i=1}^{4}[J_i]$
return FISim $(P_1^k, P_2^k)$
```

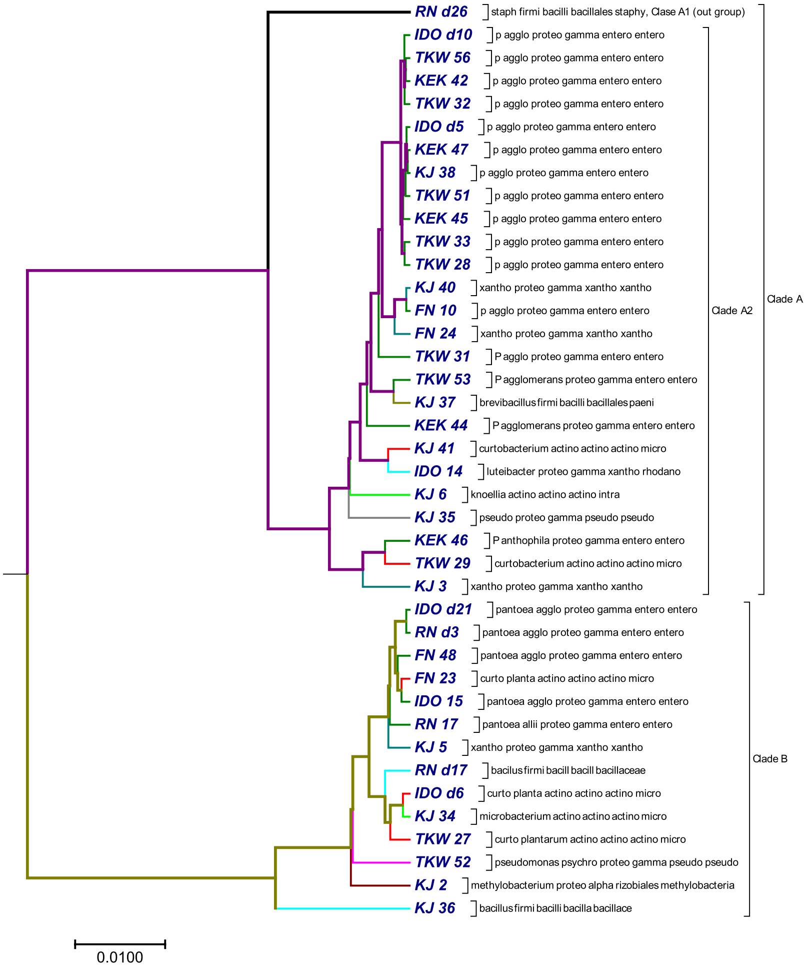

RN d26
staph firmi bacilli bacillales staphy, Clase A1 (out group)
IDO d10
p agglo proteo gamma entero entero
TKW 56
p agglo proteo gamma entero entero
KEK 42
p agglo proteo gamma entero entero
TKW 32
p agglo proteo gamma entero entero
IDO d5
p agglo proteo gamma entero entero
KEK 47
p agglo proteo gamma entero entero
KJ 38
p agglo proteo gamma entero entero
TKW 51
p agglo proteo gamma entero entero
KEK 45
p agglo proteo gamma entero entero
TKW 33
p agglo proteo gamma entero entero
TKW 28
p agglo proteo gamma entero entero
KJ 40
xantho proteo gamma xantho xantho
FN 10
p agglo proteo gamma entero entero
FN 24
xantho proteo gamma xantho xantho
TKW 31
P agglo proteo gamma entero entero
TKW 53
P agglomerans proteo gamma entero entero
KJ 37
brevibacillus firmi bacilli bacillales paeni
KEK 44
P agglomerans proteo gamma entero entero
KJ 41
curtobacterium actino actino actino micro
IDO 14
luteibacter proteo gamma xantho rhodano
KJ 6
knoellia actino actino actino intra
KJ 35
pseudo proteo gamma pseudo pseudo
KEK 46
P anthophila proteo gamma entero entero
TKW 29
curtobacterium actino actino actino micro
KJ 3
xantho proteo gamma xantho xantho
Clade A2
Clade A
IDO d21
pantoea agglo proteo gamma entero entero
RN d3
pantoea agglo proteo gamma entero entero
FN 48
pantoea agglo proteo gamma entero entero
FN 23
curto planta actino actino actino micro
IDO 15
pantoea agglo proteo gamma entero entero
RN 17
pantoea allii proteo gamma entero entero
KJ 5
xantho proteo gamma xantho xantho
RN d17
bacilus firmi bacill bacill bacillaceae
IDO d6
curto planta actino actino actino micro
KJ 34
microbacterium actino actino actino micro
TKW 27
curto plantarum actino actino actino micro
TKW 52
pseudomonas psychro proteo gamma pseudo pseudo
KJ 2
methylobacterium proteo alpha rizobiales methylobacteria
KJ 36
bacillus firmi bacilli bacilla bacillace
Clade B
0.0100

Figure 2. The phylogenetic tree constructed by our method using 16S rDNA sequences from 40 bacterial isolates.

Stage 3: We integrate stage(1) and stage(2) for phylogenetic construction:

Algorithm 3. Construction of distance matrix.

```
1. repeat $(k = 1; k++)$
2. $P_i^k \leftarrow$ follow stage 1 for 'n' DNA sequences, where
   $i \in \{1, 2, 3...n\}$
3. for $(i = 1$ to $i = n - 1$ $step)$
4. for $(j = i + 1$ to $j = n$ $step)$ do
5. $d^k[i][j] \leftarrow d^k[j][i] \leftarrow 1 - (FISim(P_i^k, P_j^k))$, by stage 2
6. $d^k[i][i] \leftarrow 0$
7. end for
8. until root mean square distance$(d^k[i][j], d^{k+1}[i][j]) \geqslant 0.0000005$
9. return distance matrix $d^k[i][j]$ for phylogenetic construction using the PHYLIP package
```

Time complexity of the proposed algorithm. To determine the time complexity of a given algorithm, we assume that all operations took the same unit of time. The whole computational process consists of three stages. In the first stage, we calculate the transition probability matrix from the raw DNA sequences. The time complexity of stage (1) is $O(m^3nk + nl)$, where l is the average length of the DNA sequences, n is the total number of DNA sequences, m is the number of bases and k is the k^{th}-step transition probability matrix. In the second stage, we calculate the fuzzy integral similarity between the two $m \times m$ transition probability matrices generated in stage (1). The time complexity of stage (2) is $O(m2^m)$. In the third stage, we integrate stage (1) and stage (2) to generate a distance matrix. Let $k = h$ be an optimal step that satisfies condition (8) in algorithm (3). Therefore, the total time complexity to generate the final distance matrix at the h^{th} optimal step is:

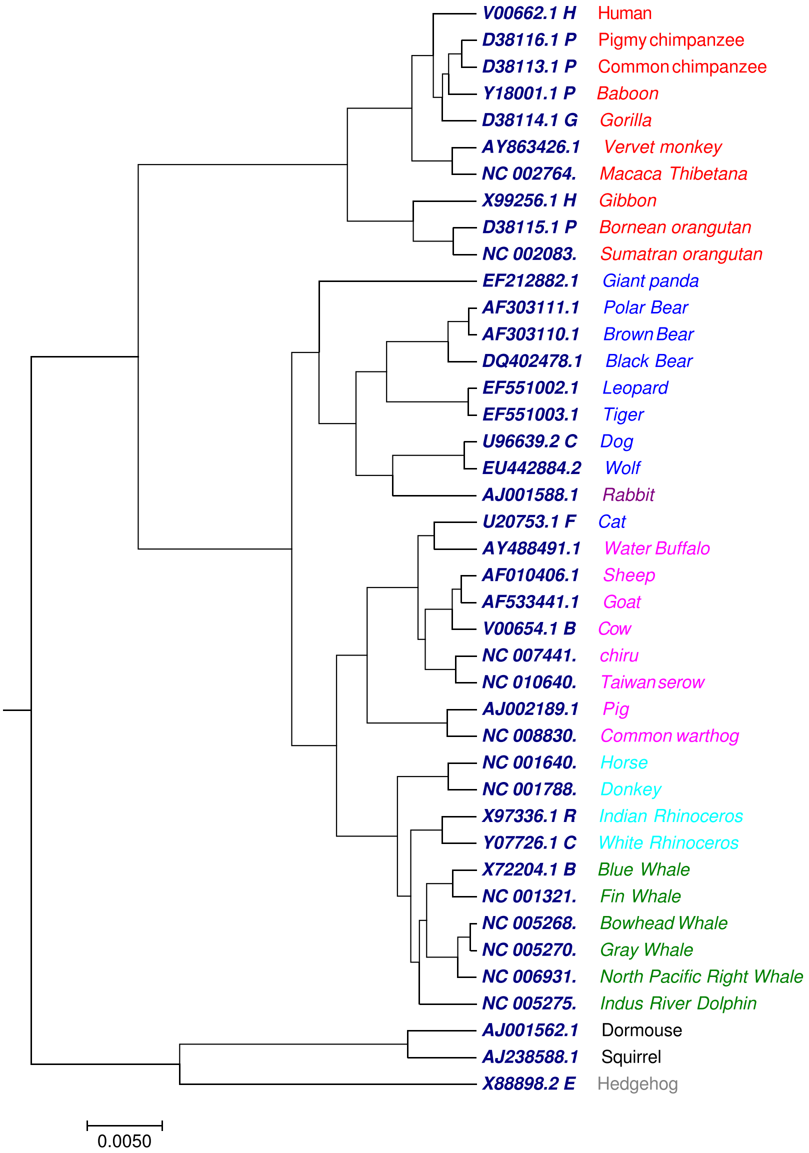


Figure 3. The phylogenetic tree of 41 mammalian mitochondrial genomes constructed using our method.

$$= h^{th} - step \text{ time complexity of stage 1} + h((n(n-1))/2) * \text{time complexity of stage 2}$$

$$= O(m^3nh + nl) + h((n(n-1))/2)_{*}O(m2^m)$$

$$= O(m^3nh + nl) + O(hn^2m2^m)$$

$$= O(m^3nh + nl + hn^2m2^m).$$

Since we are calculating the computational complexity for DNA sequences, the number of bases (A, T, G and C) is $m = 4$. Hence, the time complexity of our proposed algorithm are $O(nh + nl + hn^2)$.

Results

To check the performance of the proposed method, it was tested on different datasets. Some datasets are small sized and others are medium sized. The length of sequences ranges from seven thousands to several millions base pairs. In order to compare and analyze various genomic data, we generated a distance matrix using Eq. (12) for each distinct step k using the method described above. We increased step k until we obtained the same distance matrix for two consecutive distinct k(suppose $k = h$ *and* $h + 1$, where h is a fixed integer), (i.e., the root mean square distance[47] between two distance matrices generated by step h and $h + 1$ should be zero). Therefore, we considered $k = h$ an optimal step and generate the phylogenetic tree at step $k = h$ using the PHYLIP package. Here, we use the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) approach in the PHYLIP package[38] to generate the phylogenetic tree. To test the effectiveness of our proposed approach, we selected ten sets of test data: (i) 18S rDNA sequences from 11 Arbuscular mycorrhizal fungi isolates, (ii) 16S rDNA sequences from 40 bacterial isolates, (iii) 41 mammalian mitochondrial genomes, (iv) 59 ebolavirus complete genomes, (v) 30 coronavirus whole genomes, (vi) 30 bacterial whole genomes, (vii) 48 Hepatitis E virus (HEV) whole genomes,

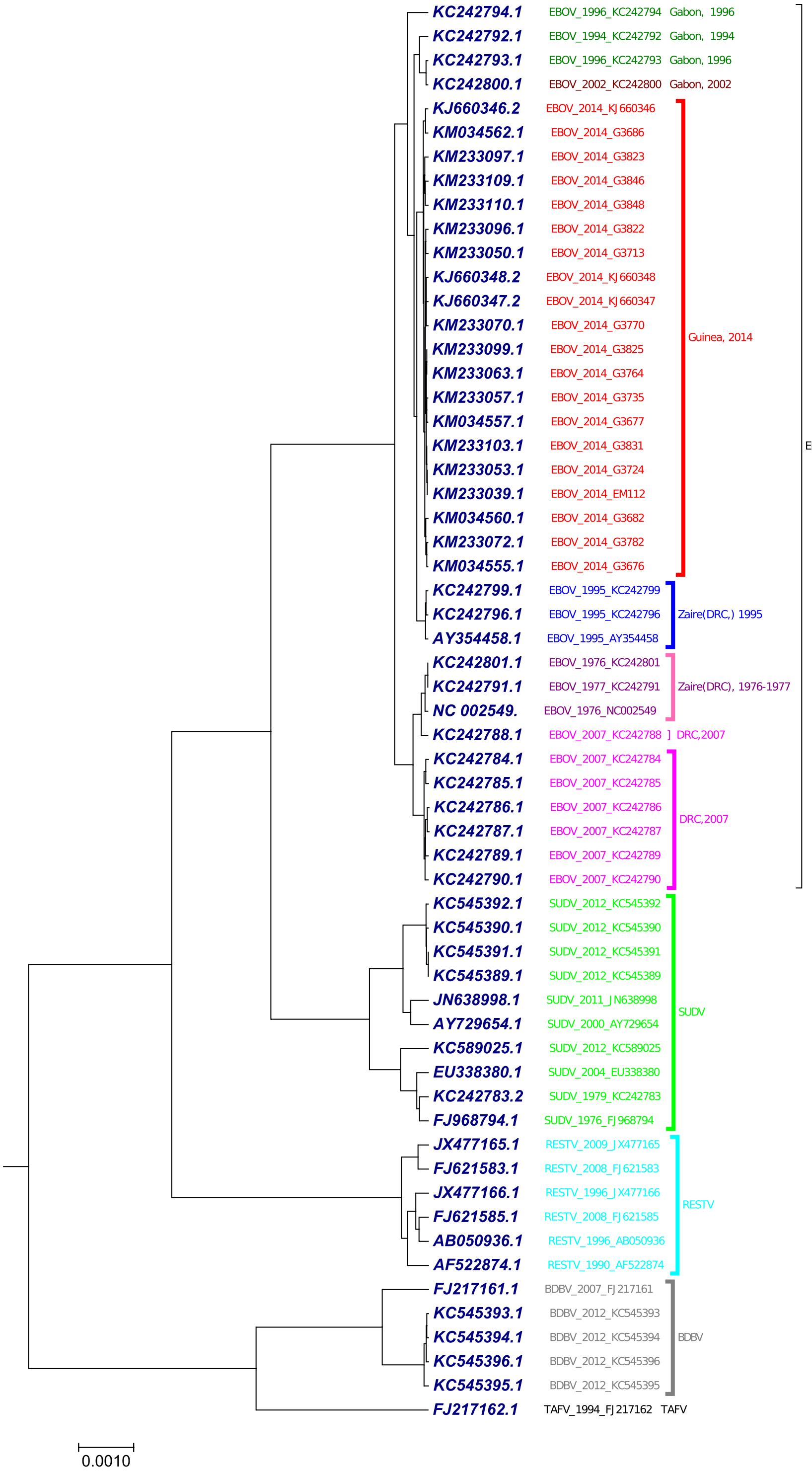

KC242794.1 EBOV_1996_KC242794 Gabon, 1996
KC242792.1 EBOV_1994_KC242792 Gabon, 1994
KC242793.1 EBOV_1996_KC242793 Gabon, 1996
KC242800.1 EBOV_2002_KC242800 Gabon, 2002
KJ660346.2 EBOV_2014_KJ660346
KM034562.1 EBOV_2014_G3686
KM233097.1 EBOV_2014_G3823
KM233109.1 EBOV_2014_G3846
KM233110.1 EBOV_2014_G3848
KM233096.1 EBOV_2014_G3822
KM233050.1 EBOV_2014_G3713
KJ660348.2 EBOV_2014_KJ660348
KJ660347.2 EBOV_2014_KJ660347
KM233070.1 EBOV_2014_G3770
KM233099.1 EBOV_2014_G3825
KM233063.1 EBOV_2014_G3764
KM233057.1 EBOV_2014_G3735
KM034557.1 EBOV_2014_G3677
KM233103.1 EBOV_2014_G3831
KM233053.1 EBOV_2014_G3724
KM233039.1 EBOV_2014_EM112
KM034560.1 EBOV_2014_G3682
KM233072.1 EBOV_2014_G3782
KM034555.1 EBOV_2014_G3676
Guinea, 2014
KC242799.1 EBOV_1995_KC242799
KC242796.1 EBOV_1995_KC242796
AY354458.1 EBOV_1995_AY354458
Zaire(DRC,) 1995
KC242801.1 EBOV_1976_KC242801
KC242791.1 EBOV_1977_KC242791
NC 002549. EBOV_1976_NC002549
Zaire(DRC), 1976-1977
KC242788.1 EBOV_2007_KC242788] DRC,2007
KC242784.1 EBOV_2007_KC242784
KC242785.1 EBOV_2007_KC242785
KC242786.1 EBOV_2007_KC242786
KC242787.1 EBOV_2007_KC242787
KC242789.1 EBOV_2007_KC242789
KC242790.1 EBOV_2007_KC242790
DRC,2007
EBOV
KC545392.1 SUDV_2012_KC545392
KC545390.1 SUDV_2012_KC545390
KC545391.1 SUDV_2012_KC545391
KC545389.1 SUDV_2012_KC545389
JN638998.1 SUDV_2011_JN638998
AY729654.1 SUDV_2000_AY729654
KC589025.1 SUDV_2012_KC589025
EU338380.1 SUDV_2004_EU338380
KC242783.2 SUDV_1979_KC242783
FJ968794.1 SUDV_1976_FJ968794
SUDV
JX477165.1 RESTV_2009_JX477165
FJ621583.1 RESTV_2008_FJ621583
JX477166.1 RESTV_1996_JX477166
FJ621585.1 RESTV_2008_FJ621585
AB050936.1 RESTV_1996_AB050936
AF522874.1 RESTV_1990_AF522874
RESTV
FJ217161.1 BDBV_2007_FJ217161
KC545393.1 BDBV_2012_KC545393
KC545394.1 BDBV_2012_KC545394
KC545396.1 BDBV_2012_KC545396
KC545395.1 BDBV_2012_KC545395
BDBV
FJ217162.1 TAFV_1994_FJ217162 TAFV
0.0010

Figure 4. The phylogenetic tree of 59 ebolavirus complete genomes constructed using our method.

(viii) 24 Eutherian mammals sequences, (ix) 58 genome datasets from different species and (x) 29 Escherichia/Shigella complete genomes. We compared our tree with the tree generated by the previously published method using same datasets.

Phylogenetic tree analysis using 18S rDNA sequences from 11 arbuscular mycorrhizal fungi (AMF) isolates.

Arbuscular mycorrhizal fungi (which is also called an AM fungi (AMF) or endomycorrhiza) is a type of mycorrhiza in which the fungus infects vascular plants by penetrating the cortical cells of the root. Arbuscular mycorrhizas are characterized by the formation of unique structures (arbuscules) and vesicles, these fungi belong to phylum *glomeromycota*. Arbuscular mycorrhizas fungi help plants to capture nutrients, such as

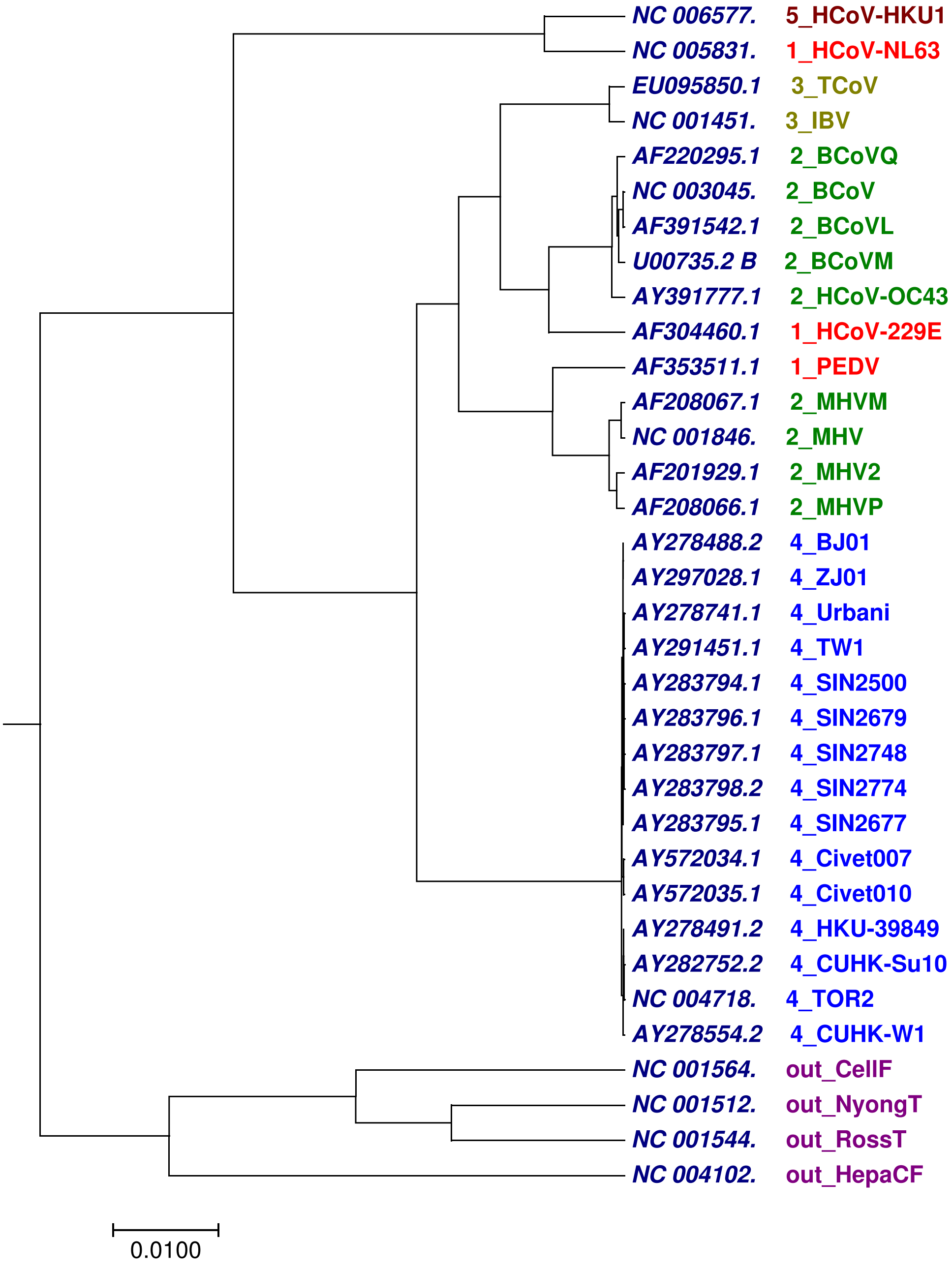


Figure 5. The phylogenetic tree of 30 coronavirus whole genomes constructed using our method.

phosphorus, sulfur, nitrogen and micronutrients, from soil. The development of arbuscular mycorrhizal symbiosis is believed to have played a crucial role in the initial colonization of plants on land and in the evolution of vascular plants[48]. We built a phylogenetic tree (Fig. 1) using the optimal step $k = 8$ of 11 AMF sequences listed in Table S1. To compare our method with an alignment-based method, we also constructed the phylogenetic tree (Fig. S1) by ClustalW method using MEGA package[49]. We characterized 11 AMF sequences based on their families and genera. All *rhizophagus* genera belonging to family *glomeraceae* were clustered together in cluster (a), except one genus of *rhizophagus* (i.e., the "15 Rhi in" sequence belongs to cluster (d)). All *glomus* genera belonging to family *glomeraceae* were clustered in cluster (b). All *gigaspora* genera belonging to family *gigasporaceae* were clustered in cluster (c). While comparing the tree prepared by our method (Fig. 1) with the tree prepared by ClustalW method (Fig. S1) using the UPGMA approach, we found that, *glomus* genera were clustered together in Fig. 1 which was lacking in Fig. S1. An obvious flaw in both the phylogenetic trees (Figs 1 and S1) is, none of them clustered *rhizophagus* genera in the single clade.

Phylogenetic tree analysis using 16S rDNA sequences from 40 bacterial isolates.

Endophytic bacteria are an essential part of plant systems and play significant roles in plant growth and development[50]. The 40 bacterial sequences were obtained from pure cultures of endophytic bacteria isolated from the

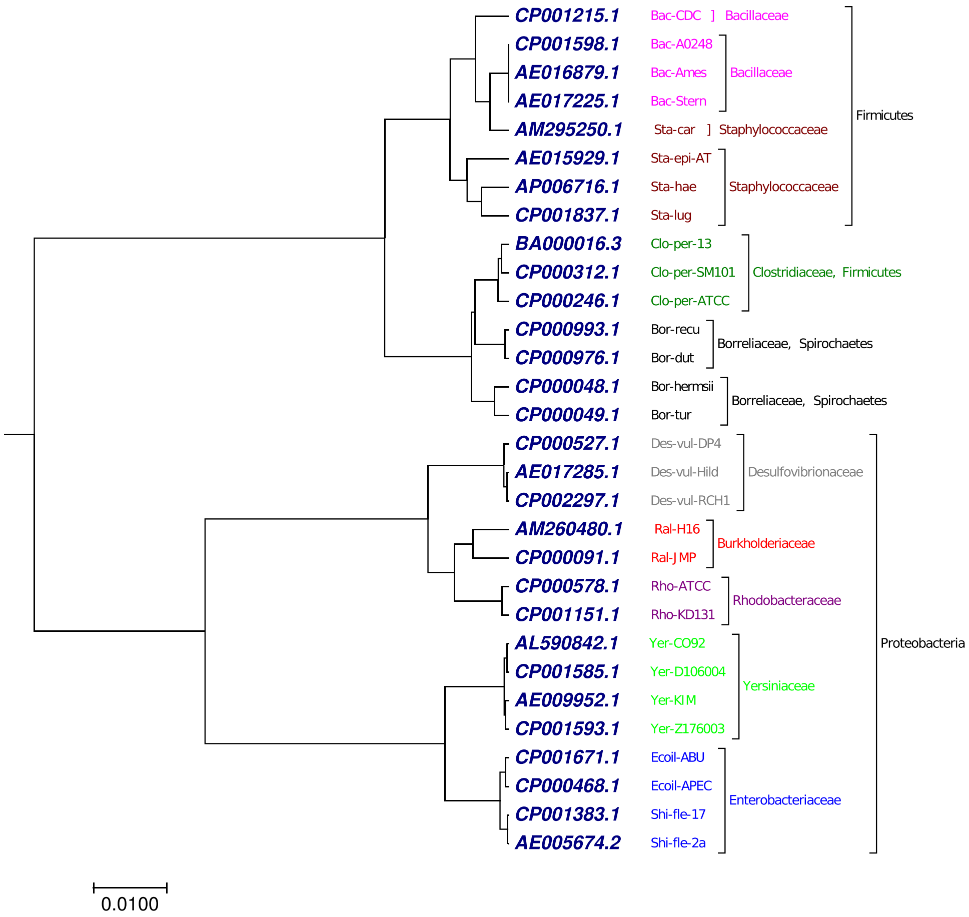


Figure 6. The phylogenetic tree of 30 bacterial whole genomes constructed using our method.

surface sterilized mature endosperms of six rice varieties. The rice seeds were collected from two different locations in north-east India: North-Lakhimpur, Assam, Aizawl and Mizoram. Genomic DNA was extracted from the pure cultures, and the full length 16S rDNA sequences were amplified using the primer pair 27f (5′-AGAGTTTGATYMTGGCTCAG) and 1492r (5′-TACCTTGTTAYGACTT). The amplicons were sequenced in an Applied Biosystems sequencer using the BigDye terminator method. To minimize sequencing error, we also used two internal primers 533f/805r along with 27 f/1492r. We assembled the contigs based on their phred scores (15) using the Codon Code aligner v7.0.1, BioEdit and SeqTrace v0.9 software[51]. The contigs were checked for the presence of any chimeras in mothur v1.35.1 and were aligned for identification against the NCBI reference rRNA database using the blastn algorithm. Information, including the accession numbers, phyla, classes, orders and families for the 40 bacterial isolates are listed in Table S2. With there 40 bacterial isolates, we generated a phylogenetic tree (Fig. 2) with our approach using the optimal step $k = 6$. The tree (Fig. 2) obtained by our method was compared with the tree (Fig. S2) obtained by ClustalW method using MEGA package[49]. Our algorithm separated the 40 bacterial sequences into two major clades: clade A (purple) and clade B (green) (Fig. 2). Clade A branched into two clades: clade A1 (bold black) and clade A2 (purple). Clade A1 contained only *Staphylococcus warneri*, which separated out as an outgroup from the sequences in clade A2. Clade A2 consisted of 25 sequences, of which one sequence represented phylum *firmicutes*, three sequences represented *actinobacteria*, and the remaining 21 sequences belonged to phylum *proteobacteria*. Our method successfully grouped sequences of genus pantoea together, but in one instance it placed pantoea and xanthomonas as sister groups. In the same clade, our method placed a third sequence belonging to xanthomonas as an outlier. Additionally, in clade A2, brevibacillus and pantoea were clustered as a sister group, which belong to phylum *firmicutes* and *proteobacteria* respectively and curtobacterium was grouped with luteibacter. None of the actinobacterial sequences were grouped together in this clade.

The second major cluster or clade B(green) consisted of 14 sequences. *Bacillus marisflavi* was an outlier from the remaining 13 sequences. In this cluster, *Curtobacterium plantarum* and *Pantoea agglomerans* were placed together as sister groups, which might indicate sequence similarity between the two species. Three actinobacterial species (two sequences of *C. plantarum* and one sequence of *Microbacterium proteolyticum*) were placed together in one clade. However, *C. plantarum* and *M. proteolyticum* were placed as sister groups, and the other *C. plantarum* sequence was positioned as an outlier. When we compared our method (Fig. 2) with ClustalW method (Fig. S2), we found that both methods clearly separated the 40 bacterial sequences into two major clades. Each

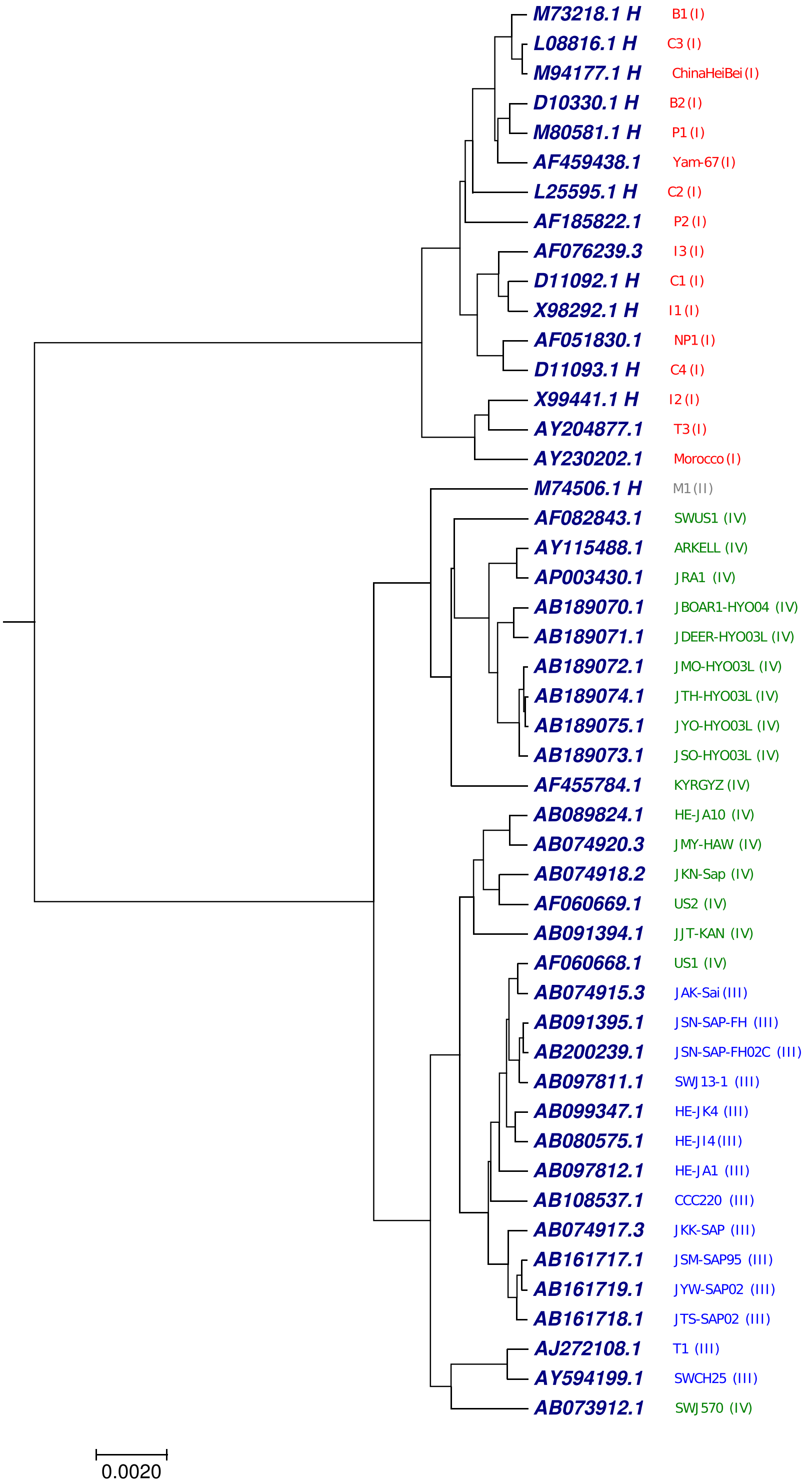


Figure 7. The phylogenetic tree of 48 Hepatitis E virus (HEV) whole genomes constructed using our method.

clade contains the same type of bacterial sequences, but the order was interchanged. In clade A, our method failed to cluster xanthomonas together, which was grouped together in the result obtained by ClustalW.

Phylogenetic tree analysis of 41 mammalian mitochondrial genomes. The proposed algorithm was tested on the benchmark mammalian dataset containing 41 complete mitochondrial genomes(mtDNA) with nearly 16500 nucleotides (Table S3). The tree generated by our approach (Fig. 3) using the optimal step $k = 6$, the 41 species were correctly divided into eight groups: *Primates* (red), *Cetacea* (green), *Artiodactyla* (pink), *Perissodactyla* (light green), *Rodentia* (black), *Lagomorpha* (dark red), *Carnivore* (blue), and *Erinaceomorpha* (grey). The cat species in our approach was clustered with the *Artiodactyla* group. We compared the phylogenetic

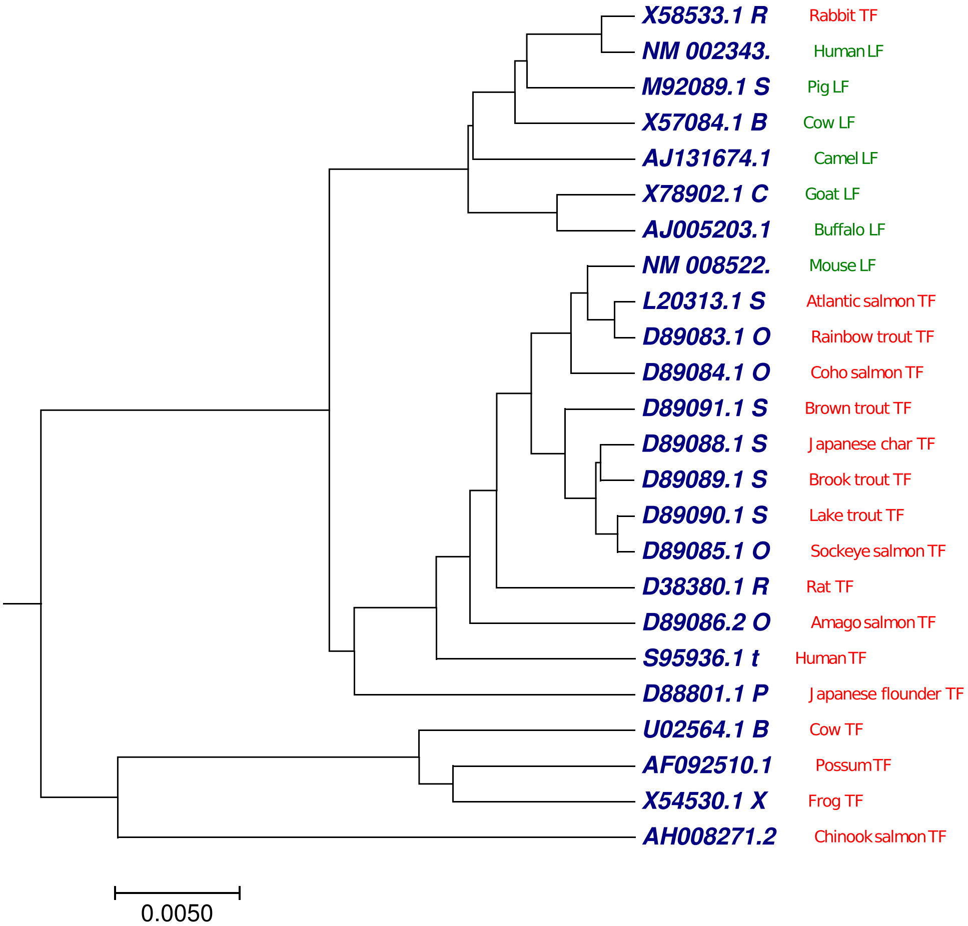


Figure 8. The phylogenetic tree of 24 Eutherian mammals sequences constructed using our method.

tree (Fig. 3) generated by our approach with the phylogenetic tree (Figs S3 and S4) collected from previous work[52]. The Fig. S3 is generated by multiple encoding vector method[52] and Fig. S4 is generated by FFP method[22] using substrings length seven. In Fig. 3, the 10 primates (red) formed a cluster, also Vervet monkey and Macaca Thibetana of family *cercopithecidae* were clustered in a single clade as sister group which was not observed in Fig. S3. Moreover, species belong to *Artiodactyla* were grouped into a separate clade, which was lacking in Fig. S3. We have also compared our result with the phylogenetic tree (Fig. S4) generated by FFP method. As showed in Fig. S4, the eight groups were not classified well. The four species of *Perissodactyla* were distributed into two clades. Indus RiverDolphin from *Cetacea* was separated from other species of *Cetacea*. The *Primates*, *Artiodactyla* and *Carnivore* clades were all divided into more than one group. The phylogenetic tree (Fig. 3) generated by our approach shows a better clustering as compared to Figs S3 and S4.

Phylogenetic tree analysis of 59 ebolavirus complete genomes. The benchmark dataset used in this study was 59 complete genomes of ebolavirus with nearly 18900 nucleotides (Table S4). The *Ebolavirus* genus includes five species: Bundibugyo virus (BDBV), Reston virus (RESTV), Ebola virus (formerly Zaire ebolavirus, EBOV), Sudan virus (SUDV), and Tai Forest virus (TAFV)[53]. Ebola viruses are single-strand negative sense RNA viruses. Each ebolavirus genome encodes seven proteins in which glycoprotein is the only viral protein on the surface of ebolavirus. The first case of human, infected by EBOV, was reported in 1976 in Zaire (currently the Democratic Republic of the Congo (DRC))[54]. We applied our proposed method to generate the phylogenetic tree (Fig. 4) using the optimal step $k = 6$ of 59 viruses in *Ebolavirus* genus. As shown in Fig. 4, the five species were correctly separated. The EBOV strains from the recognized pandemics build a lineage independent of the other four species in genus *Ebolavirus*. The EBOV strains in Zaire (DRC) pandemic in 1976–1977 were clustered together as a clade. The EBOV strains in DRC pandemic in 2007 were clustered together with the exception *EBOV*_2007_*KC*242788, which was clustered with Zaire (DRC) in 1976–1977. The EBOV strains in Guinea epidemic in 2014 were clustered together as a clade. The three EBOV strains from the 1995 outbreak in Zaire (DRC) formed a clade. SUDV and RESTV formed separate clades. BDBV and TAFV viruses were positioned together. Our result was in consensus with the result generated using multiple encoding vector method[52] (Fig. S5) and FFP method[22] (Fig. S6).

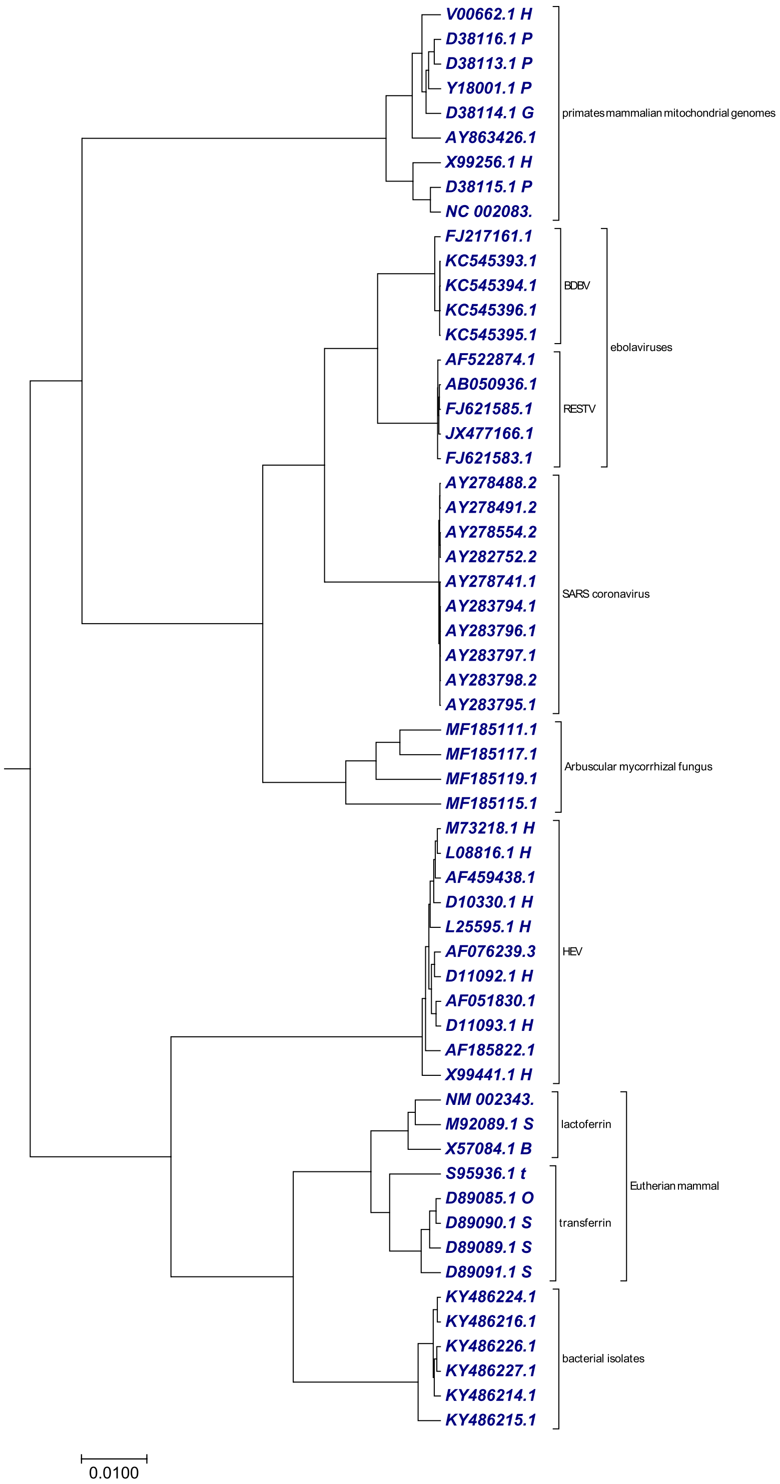


Figure 9. The phylogenetic tree of 58 genome datasets from different species constructed using our method.

Phylogenetic tree analysis of 30 coronavirus whole genomes. The other benchmark dataset used for validation of the method was the 30 complete coronavirus genomes with nearly 25,000 to 32,000 nucleotides (Table S5). Coronaviruses[55] are enveloped, single-stranded, positive-sense RNA viruses within the family *Coronaviridae*[56]. The coronaviruses are pleomorphic RNA viruses that are widespread among avians, bats, humans and other mammals. They are known to cause mild to severe respiratory diseases, gastroenterological, neurological and systemic conditions. This group of virus can easily cross species-barrier and infect new species[57]. As a result of pandemics from coronaviruses especially the SARS, the classification and evolutionary relationships among these viruses have been extensively investigated. We employed our method to analyse the 30 coronavirus whole genome sequences along with 4 non-coronaviruses as outgroups. The 30 coronavirus were classified into

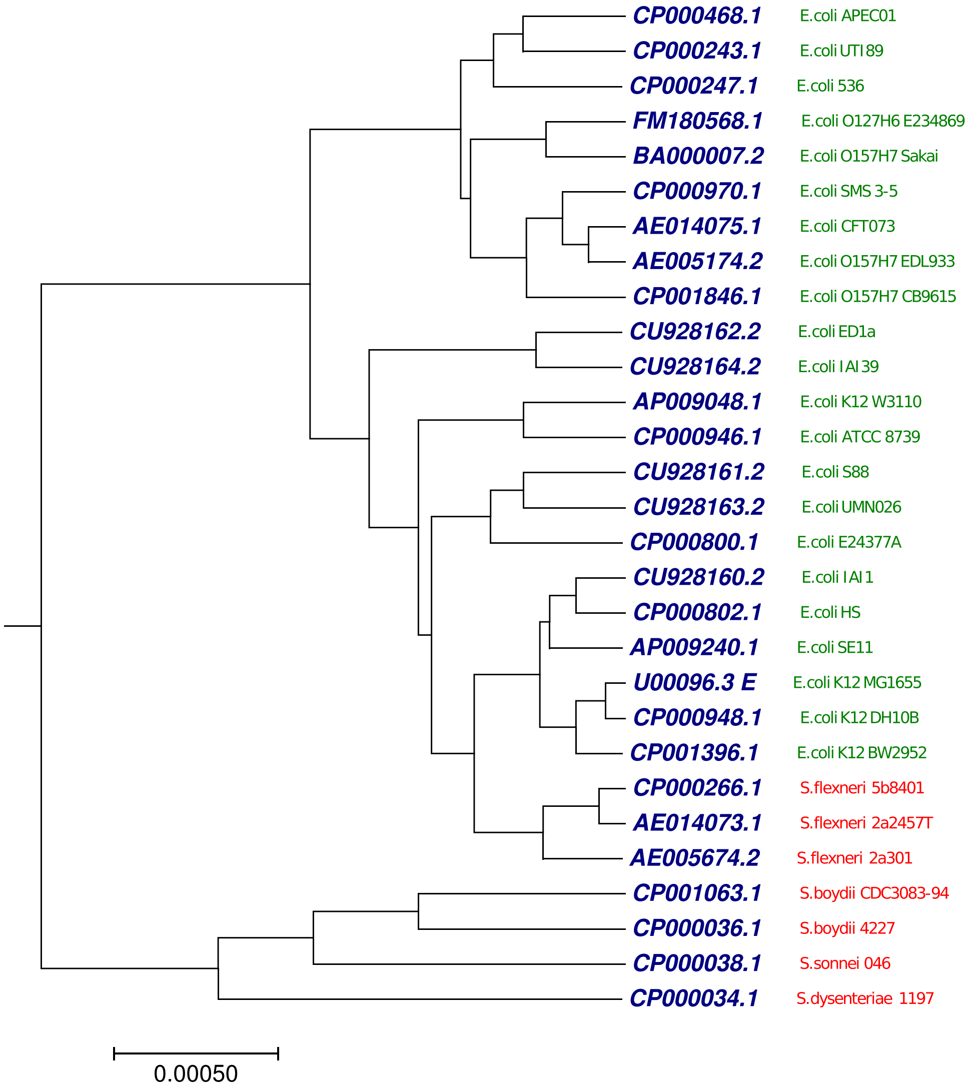

CP000468.1 E.coli APEC01
CP000243.1 E.coli UTI89
CP000247.1 E.coli 536
FM180568.1 E.coli O127H6 E234869
BA000007.2 E.coli O157H7 Sakai
CP000970.1 E.coli SMS 3-5
AE014075.1 E.coli CFT073
AE005174.2 E.coli O157H7 EDL933
CP001846.1 E.coli O157H7 CB9615
CU928162.2 E.coli ED1a
CU928164.2 E.coli IAI39
AP009048.1 E.coli K12 W3110
CP000946.1 E.coli ATCC 8739
CU928161.2 E.coli S88
CU928163.2 E.coli UMN026
CP000800.1 E.coli E24377A
CU928160.2 E.coli IAI1
CP000802.1 E.coli HS
AP009240.1 E.coli SE11
U00096.3 E E.coli K12 MG1655
CP000948.1 E.coli K12 DH10B
CP001396.1 E.coli K12 BW2952
CP000266.1 S.flexneri 5b8401
AE014073.1 S.flexneri 2a2457T
AE005674.2 S.flexneri 2a301
CP001063.1 S.boydii CDC3083-94
CP000036.1 S.boydii 4227
CP000038.1 S.sonnei 046
CP000034.1 S.dysenteriae 1197
0.00050

Figure 10. The phylogenetic tree of 29 Escherichia/Shigella whole genomes constructed using our method.

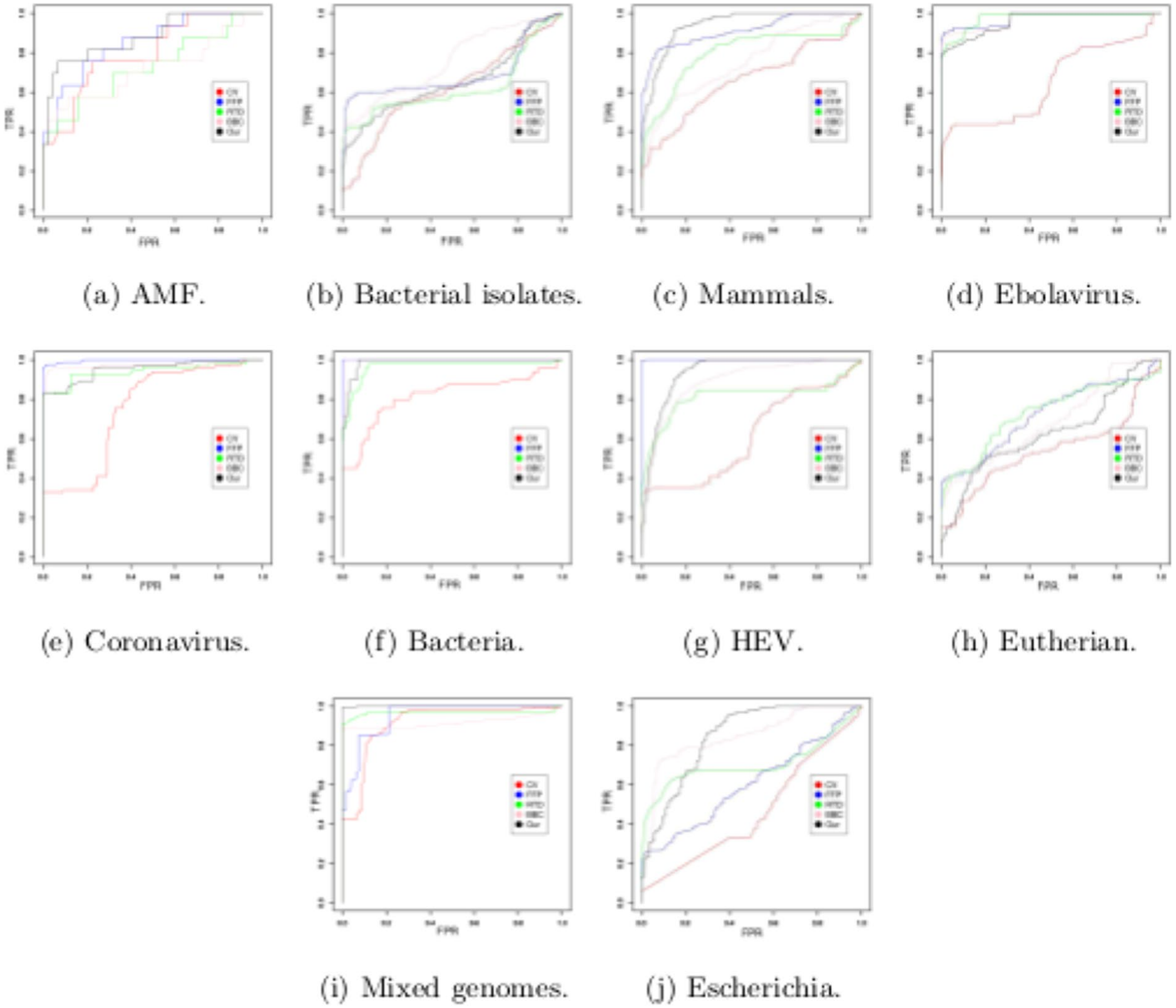
five groups according to their host type. As shown in Fig. 5 generated by our approach using the optimal step $k = 7$, we can observe that the 30 coronavirus along with 4 non-coronaviruses were correctly grouped according to their host type except group 1 (Table S5). We compared Fig. 5 generated by our approach with Figs S7, S8, S9 and S10 collected from previous work[52,58]. The limitation observed in our result (Fig. 5) is that, our method was unable to cluster group 1 as compared to Figs S7 and S8. While in Fig. S9 generated by k–mer[59] method and Fig. S10 generated by FFP method using substrings length six, the four non-coronaviruses were not clustered together. Therefore, for this dataset, tree generated by our approach has advantage over k–mer and FFP methods using substrings length six.

Phylogenetic tree analysis of 30 bacterial whole genomes. Another benchmark dataset used in this study was 30 complete bacterial genomes with more than 1 million nucleotides (Table S6). Methods based on multiple sequence alignment program cannot handle such large dataset. As shown in Fig. 6, generated by our approach using the optimal step $k = 7$, the 30 bacterial genomes were clustered into nine groups based on taxnomic family: *Burkholderiaceae*, *Rhodobacteriaceae*, *Enterobacteriaceae*, *Borreliaceae*, *Bacilleceae*, *Clostridiaceae*, *Desulfovibrionaceae*, *Yersiniaceae*, and *Staphylococcaceae*. Our result (Fig. 6) has similarity with the result (Fig. S11) generated by fourier power spectrum method at the taxnomic family level collected from previous work[58]. However, our phylogenetic tree (Fig. 6) has advantages at the phylum level which was lacking in Fig. S11. As shown in Fig. 6, the genomes were successfully clustered into three phylum, *Firmicutes*, *Proteobacteria*, and *Spirochaetes* as a separate clade, which was not observed in Fig. S11.

Phylogenetic tree analysis of 48 Hepatitis E virus (HEV) whole genomes. The other benchmark was 48 complete genomes of hepatitis E virus (HEV). This virus is characterized as non-enveloped, single-stranded RNA virus with nearly 7200 nucleotides (Table S7). The acute condition of the disease is caused

Method	AMF isolates	Bacterial isolates	Mammals	Ebolavirus	Coronavirus	Bacteria	HEV	Eutherian mammals	Mixed genomes	Escherichia/Shigella
Our method	<1 s	<1 s	<1 s	<1 s	<1 s	3 s	<1 s	<1 s	<1 s	3 s
ClustalW	1 s	1 min 8 s	4 h 75 min	10 h 28 min	5 h 23 min	—	1 h 28 min	2 min 13 s	3 h 12 m	—
Multiple encoding vector method	—	—	0.12 s	6 min 42 s	0.34 s	—	—	—	—	—
Fourier power spectrum method	—	—	—	—	6 s	9 min 41 s	—	—	—	—

Table 1. Running time comparison.



(a) AMF.

(b) Bacterial isolates.

(c) Mammals.

(d) Ebolavirus.

(e) Coronavirus.

(f) Bacteria.

(g) HEV.

(h) Eutherian.

(i) Mixed genomes.

(j) Escherichia.

Figure 11. Receiver operating characteristic curve (ROC) on the given datasets.

by the hepatitis E virus. The difference between other known hepatitis viruses (A, B, C, D) and hepatitis E virus is that, the hepatitis E virus is the only animal-host disease hepatitis[60]. To understand the relationship between HEV sequences, we have applied our proposed method to generate the phylogenetic tree using the optimal step $k = 6$ of the 48 HEV whole genome sequences. As shown in Fig. 7 generated by our approach, the HEV genomes were divided into separate clades based on four genotypic[61] category (I(red), II(grey), III(blue) and IV(green)) except few sequences. Phylogenetic tree (Fig. 7) generated by our approach shows better clade distribution based on the genotypic division as compared with Figs S12 and S13 collected from the previous work[62].

Phylogenetic tree analysis of 24 Eutherian mammal sequences. We selected transferrin (red) and lactoferrin (green) sequences from 24 vertebrates as a benchmark dataset[63] (Table S8). Vertebrate transferrins and lactoferrins are iron-binding proteins found in blood serum, milk, egg whites, tears, and interstitial spaces. They can be involved in iron storage and resistance to bacterial disease. We have applied our proposed method to generate the phylogenetic tree (Fig. 8) using the optimal step $k=8$ of the 24 Eutherian mammal sequences. As shown in Fig. 8, we can observe that all transferrin sequences (red) were clustered into two distinct clades, except rabbit transferrin sequence was grouped with lactoferrin class. Similarly, all lactoferrin sequences (green) were clustered together, except mouse lactoferrin sequence was grouped with transferrin class. Phylogenetic tree (Fig. 8) generated by our approach showed better clade distribution based on transferrin and lactoferrin categories compared with previous work[62] which is shown in Figs S14 and S15.

Phylogenetic tree analysis using 58 genome datasets from different species. To verify the clustering efficiency of our method on extremely divergent sequences from different organisms, we randomly collected genomes of varying length from different datasets from Tables S1, S2, S3, S4, S5, S7 and S8. The genomes included in this dataset were, four arbuscular mycorrhizal fungi, six bacterial isolates, nine primates mammalian mitochondrial genomes, ten ebolavirus (five reston virus (RESTV), five bundibugyo virus (BDBV)) complete genomes, ten SARS coronavirus, eleven hepatitis E virus and eight eutherian mammals. We applied our proposed method to generate the phylogenetic tree (Fig. 9) using the optimal step $k=8$. As shown in Fig. 9, we observed that all the different species genomes were clustered separately. This result (Fig. 9) showed the efficiency of our method in clustering genomes irrespective of their size and divergence. Our result (Fig. 9) was in consensus with the result generated using ClustalW method (Fig. S16). The time taken by our method to generate the transition probability matrix was less than 1 second, while Clustalw has taken 3 hours and 12 minutes.

Phylogenetic tree analysis using 29 Escherichia/Shigella complete genomes. The other benchmark dataset used in this study was 29 complete genomes from the genera Escherichia/Shigella with more than 1 million nucleotides (Table S9). We applied our proposed method to generate the phylogenetic tree (Fig. 10) using the optimal step $k=7$. As shown in Fig. 10, we observed that the genomes were clustered into distinct clades, Escherichia(green) and Shigella (red). We took the benchmark tree[64] as a reference which is based on concatenated alignments of the 2034 core genes and used the maximum likelihood method to infer the phylogenetic relationships. We calculated Robinson-Foulds distance (RF-distance)[65] of the tree produced by our method against the benchmark tree[64]. The RF-distance is often used to compare two trees of closely related species. Since, the species in this dataset (29 complete genomes from the genera Escherichia/Shigella) are closely related organisms, therefore, we employed the RF-distance, which evaluates the topological congruence between an inferred tree and a benchmark tree. We also collected the generated RF-distances from the previous study[66]. $\mathrm{RF}=0$ indicates that, the test-tree topology is completely similar to that of the benchmark tree, while similarity level decreases as the RF increases. As shown in Fig. S17, RF-distance generated by our approach to the reference tree was higher than RF-distance generated by rest of the methods to the reference tree. This result demonstrates that our proposed method has a limitation in clustering of the closely related organism.

Conclusion

This study focused on fuzzy integral similarity technique based on Markov chain and applied this algorithm to phylogenetic tree analysis. Sequence comparison is one of the most useful and widely practiced methods in bioinformatics and computational biology. Alignment based methods perform well if the genetic sequences are homologous. High mutation rates and genetic recombination brings in a limitation of the alignment based method. Also at the genomic scale, alignment based methods become impractical due to their computational complexity. Alignment-free methods are of great value, because they reduce the technical constraints of alignments. We constructed a transition probability matrix using a Markov chain for each DNA sequences without performing prior alignment at the genomic scale. The fuzzy integral similarity technique is a method that can calculate similarity score between two transition probability matrices of DNA sequences. The main advantage of our approach is that, it does not require any additional parameters, which makes it fully automated and robust. We implemented and tested our method on suitable datasets.

All programs are implemented on a linux server with 384 GB RAM with 24 dual core processor. Our proposed approach is fast in computational speed (Table 1) compare to alignment-based method, ClustalW and also faster as compared to various alignment-free methods, which were discussed above. For the large datasets such as, 30 bacteria and 29 Escherichia/Shigella, which ClustalW can not handle, while our alignment-free method take only 3 seconds to produce transition probability matrices for both the datasets.

In this study, we plotted ROC curve[39–41] (Fig. 11) and calculate area under the ROC curve (AUC) using distance matrices generated by our method and other alignment-free methods from Alfree repository[2]. The detail discussion of the ROC results (Fig. 11) and AUC analysis for all benchmark datasets are given in ROC_Supplementary Material. It may be observed that, while we have similar AUC values as the other methods, the phylogenetic tree generated by our method outperforms the other existing methods. The result shows clear accuracy in terms of AUC of our method as the other methods and superiority in terms of phylogenetic clustering. Moreover, the superiority of our method can be observed from the execution time in Figs 30 and 48 (ROC_Supplementary Material) for the large sequence length data.

Our proposed method is faster and has the potential to build phylogenetic tree for large sized genomes, such as, mammalian genome. Mammalian genomes are divided into several chromosomes. In this study, we selected chromosome X to do the phylogenetic analysis, details are given in Table S10. Our dataset included the species: human (*Homo sapiens*), monkey (*Macaca mulatta*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), horse (*Equus caballus*), mouse (*Mus musculus*), dog (*Canis familiaris*), opossum (*Monodelphis domesticus*), and platypus (*Ornithorhynchus anatinus*). The length of the chromosomes X in these organisms ranges approximately from 6 to 147 Mb. Applying our method, we generated the phylogenetic tree (Fig. S18) of nine mammals using the optimal step $k = 8$. Our method took only 18 seconds for generating transition probability matrix. Phylogenetic tree (Fig. S19) generated by multiple encoding vector method[52], mouse clustered with primates, while dog and horse came together in a clade. Figure S18 generated by our method formed three major clusters. Branch point in the first clade shows a divergence event of horse and mouse from the primates. A comparative radiation hybrid map of chromosome X of human, horse and mouse reveals many conserved syntenies between the three species[67,68]. This similarity may have placed horse and mouse in a sister group. Dog and oppossum formed a distinct clade, while platypus separated as an outgroup.

Based on the results generated by our developed method, we found that our method performed well on divergent sequences, rather than closely related sequences. Therefore, this approach would be beneficial for the users to generate hypothesis that can be investigated in further detail with subsequent analysis. Before continuing research for further development of our method, we must keep in mind that, this approach is a probabilistic measure in nature, and can be modified by incorporating more information, such as, nucleotides substitution, insertion, deletion, genetic recombination, Physicochemical Properties of nucleotides etc., in sequences. Overall, our goal in this study was to introduce a new methodology and a new tool to the comparative genomics research community. This proposed work can be used to guide the development of more powerful measures for sequence comparison.

Data Availability

The datasets used in this paper are available in the supplementary table and the C source code in this paper is freely available to the public upon request.

References

1. Vinga, S. & Almeida, J. Alignment-free sequence comparison—a review. *Bioinforma.* **19**, 513–523 (2003).
2. Zielezinski, A., Vinga, S., Almeida, J. & Karlowski, W. M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* **18**, 186 (2017).
3. Bernard, G. *et al.* Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings Bioinforma.* bbx067 (2017).
4. Bromberg, R., Grishin, N. V. & Otwinowski, Z. Phylogeny reconstruction with alignment-free method that corrects for horizontal gene transfer. *PLOS Comput. Biol.* **12**, 1–39 (2016).
5. Didier, G. *et al.* Comparing sequences without using alignments: application to hiv/siv subtyping. *BMC Bioinforma.* **8**, 1 (2007).
6. Chatterji, S., Yamazaki, I., Bai, Z. & Eisen, J. A. *Compostbin: A dna composition-based algorithm for binning environmental shotgun reads*. 17–28 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).
7. Meinicke, P. Uproc: tools for ultra-fast protein domain classification. *Bioinforma.* **31**, 1382–1388 (2015).
8. Tanaseichuk, O., Borneman, J. & Jiang, T. Separating metagenomic short reads into genomes via clustering. *Algorithms for Mol. Biol.* **7**, 27 (2012).
9. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glöckner, F. O. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinforma.* **5**, 163 (2004).
10. Wang, Y., Leung, H. C., Yiu, S. & Chin, F. Y. Metacluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinforma.* **28**, i356–i362 (2012).
11. Wu, Y.-W. & Ye, Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J. Comput. Biol.* **18**, 523–534 (2011).
12. Federico, M., Leoncini, M., Montangero, M. & Valente, P. Direct vs 2-stage approaches to structured motif finding. *Algorithms for Mol. Biol.* **7**, 20 (2012).
13. Kantorovitz, M. R., Robinson, G. E. & Sinha, S. A statistical method for alignment-free comparison of regulatory sequences. *Bioinforma.* **23**, i249–i255 (2007).
14. Leung, G. & Eisen, M. B. Identifying cis-regulatory sequences by word profile similarity. *Plos One* **4**, 1–11 (2009).
15. Lingner, T. & Meinicke, P. Remote homology detection based on oligomer distances. *Bioinforma.* **22**, 2224–2231 (2006).
16. Lingner, T. & Meinicke, P. Word correlation matrices for protein sequence analysis and remote homology detection. *BMC Bioinforma.* **9**, 259 (2008).
17. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
18. Rob Patro, S. M. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32**, 462–464 (2014).
19. Drouin, A. *et al.* Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics* **17**, 754 (2016).
20. Haubold, B. Alignment-free phylogenetics and population genetics. *Briefings Bioinforma.* **15**, 407–418 (2014).
21. Blaisdell, B. E. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a variety of computer-generated model systems. *J. Mol. Evol.* **32**, 521–528 (1991).
22. Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H. Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proc. Natl. Acad. Sci.* **106**, 2677–2682 (2009).
23. Kolekar, P., Kale, M. & Kulkarni-Kale, U. Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping. *Mol. Phylogenetics Evol.* **65**, 510–522 (2012).
24. Hatje, K. & Kollmar, M. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. *Front Plant Sci.* **3** (2012).
25. Lu, G., Zhang, S. & Fang, X. An improved string composition method for sequence comparison. *BMC Bioinforma.* **9**, S15 (2008).
26. Gao, L. & Qi, J. Whole genome molecular phylogeny of large dsdna viruses using composition vector method. *BMC Evol. Biol.* **7**, 41 (2007).
27. Wu, X., Wan, X.-F., Wu, G., Xu, D. & Lin, G. Phylogenetic analysis using complete signature information of whole genomes and clustered neighbour-joining method. *Int. J. Bioinforma. Res. Appl.* **2**, 219–248 (2006).
28. Ulitsky, I., Burstein, D., Tuller, T. & Chor, B. The average common substring approach to phylogenomic reconstruction. *J. Comput. Biol.* **13**, 336–350 (2006).
29. Comin, M. & Verzotto, D. Alignment-free phylogeny of whole genomes using underlying subwords. *Algorithms for Mol. Biol.* **7**, 34 (2012).
30. Haubold, B., Pierstorff, N., Möller, F. & Wiehe, T. Genome comparison without alignment using shortest unique substrings. *BMC Bioinforma.* **6**, 123 (2005).
31. Thankachan, S. V., Chockalingam, S. P., Yongchao, L., Alberto, A. & Srinivas, A. Alfred: A practical method for alignment-free distance computation. *J. Comput. Biol.* **23**, 452–460 (2016).
32. Leimeister, C.-A. & Morgenstern, B. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinforma.* **30**, 2000–2008 (2014).
33. Torra, V. & Narukawa, Y. The interpretation of fuzzy integrals and their application to fuzzy systems. *Int. J. Approx. Reason.* **41**, 43–58 (2006).

34. Medhi, J. *Stochastic Processes* (New Age Science, 2009).
35. Garcia, F., Lopez, F. J., Cano, C. & Blanco, A. Fisim: A new similarity measure between transcription factor binding sites based on the fuzzy integral. *BMC Bioinforma.* **10**, 224 (2009).
36. Zhang, S., Zhang, Y. & Gutman, I. Analysis of dna sequences based on the fuzzy integral. *Match Commun. Math. Comput. Chem.* **70**, 417–430 (2013).
37. Sims, J. R. & Zhenyuan, W. Fuzzy measures and fuzzy integrals: An overview. *Int. J. Gen. Syst.* **17**, 157–189 (1990).
38. Felsenstein, J. Phylip–phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).
39. Swets, J. Measuring the accuracy of diagnostic systems. *Sci.* **240**, 1285–1293 (1988).
40. Nemes, S. & Hartel, T. Summary measures for binary classification systems in animal ecology. *North-Western J. Zool.* **6**, 323–330 (2010).
41. Sonego, P., Kocsor, A. & Pongor, S. Roc analysis: applications to the classification of biological sequences and 3d structures. *Briefings Bioinforma.* **9**, 198–209 (2008).
42. Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological Sequence Analysis* (Cambridge University Press, Cambridge, 1998).
43. Zadeh, L. Fuzzy sets. *Inf. Control.* **8**, 338–353 (1965).
44. Sugeno, M. Theory of Fuzzy Integrals and Its Applications (Doct. Thesis, Tokyo Institute of Technology, Tokyo, 1974).
45. Sugeno, M. *Fuzzy measures and fuzzy integrals: A survey*, 89–102 (North Holland, New York, 1997).
46. Chaira, T. *Fuzzy Measures in Image Processing*, 587–606 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).
47. Carugo, O. & Pongor, S. A normalized root-mean-spuare distance for comparing protein three-dimensional structures. *Protein Sci.* **10**, 1470–1473 (2001).
48. C. Brundrett, M. Coevolution of roots and mycorrhizas of land plants. *New Phytol.* **154**, 275–304 (2002).
49. Kumar, S., Stecher, G. & Tamura, K. Mega7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
50. Bulgarelli, D., Schlaeppi, K., Spaepen, S., van Themaat, E. V. L. & Schulze-Lefert, P. Structure and functions of the bacterial microbiota of plants. *Annu. Rev. Plant Biol.* **64**, 807–838 (2013).
51. Stucky, B. J. Seqtrace: A graphical tool for rapidly processing dna sequencing chromatograms. *J. Biomol. Tech.* **23**, 90–93 (2012).
52. Li, Y., He, L., He, R. L. & Yau, S. S.-T. A novel fast vector method for genetic sequence comparison. *Sci. Reports* **7** (2017).
53. Gire, S. K. *et al.* Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Sci.* **345**, 1369–1372 (2014).
54. Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. The evolution of ebola virus: Insights from the 2013–2016 epidemic. *Nat.* **538**, 193–200 (2016).
55. Leibowitz, J. L. *Coronaviruses: Molecular and Cellular Biology*, 693–694 (Caister Academic Press, 2008).
56. King, M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. (eds). *Family - Coronaviridae*, 806–828 (Elsevier, San Diego, 2012).
57. Greenwood, D., Barer, M., Slack, R. & Irving, W. (Elsevier, Churchill Livingstone, 2012).
58. Hoang, T. *et al.* A new method to cluster dna sequences using fourier power spectrum. *J. Theor. Biol.* **372**, 135–145 (2015).
59. Yang, K. & Zhang, L. Performance comparison between k -tuple distance and four model-based distances in phylogenetic tree reconstruction. *Nucleic Acids Res.* **36**, e33 (2008).
60. Meng, X. J. Recent advances in hepatitis e virus. *J. Viral Hepat.* **17**, 153–161 (2010).
61. Li, L. *et al.* Full-genome nucleotide sequence and analysis of a chinese swine hepatitis e virus isolate of genotype 4 identified in the guangxi zhuang autonomous region: Evidence of zoonotic risk from swine to human in south china. *Liver Int.* **29**, 1230–1240 (2009).
62. Liu, L., Li, C., Bai, F., Zhao, Q. & Wang, Y. An optimization approach and its application to compare dna sequences. *J. Mol. Struct.* **1082**, 49–55 (2015).
63. Ford, M. J. Molecular evolution of transferrin: Evidence for positive selection in salmonids. *Mol. Biol. Evol.* **18**, 639–647 (2001).
64. Zhou, Z. *et al.* Derivation of *escherichia coli* o157:h7 from its o55:h7 precursor. *Plos One* **5**, 1–14 (2010).
65. Robinson, D. & Foulds, L. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
66. Morgenstern, B., Zhu, B., Horwege, S. & Leimeister, C. A. Estimating evolutionary distances between genomic sequences from spaced-word matches. *Algorithms for Mol. Biol.* **10**, 5 (2015).
67. Chowdhary, B. P. *et al.* The first-generation whole-genome radiation hybrid map in the horse identifies conserved segments in human and mouse genomes. *Genome Res.* **13**, 742–751 (2003).
68. Raudsepp, T. *et al.* Exceptional conservation of horse–human gene order on x chromosome revealed by high-resolution radiation hybrid mapping. *Proc. Natl. Acad. Sci.* **101**, 2386–2391 (2004).

Acknowledgements

Ramalingaswami Fellowship from Department of Biotechnology, Ministry of Science and Technology, Government of India, supported S.N. (BT/RLF/Re-entry/48/2013).

Author Contributions

A.S. and B.T. developed the idea, A.S. and S.N. written code in programming language and analyzed the results, and A.S. wrote the manuscript text. B.T. and S.N. guided the study. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-40452-6.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.