

SOFTWARE

Open Access



TreeWave: command line tool for alignment-free phylogeny reconstruction based on graphical representation of DNA sequences and genomic signal processing

Nasma Boumajdi¹, Houda Bendani¹, Lahcen Belyamani^{2,3,4} and Azeddine Ibrahimi^{1*} 

*Correspondence:
a.ibrahimi@um5r.ac.ma

¹ Laboratory of Biotechnology (MedBiotech), Rabat Medical & Pharmacy School, Bioinova Research Center, Mohammed V University in Rabat, Rabat, Morocco

² Mohammed VI Center for Research and Innovation (CM6), Rabat, Morocco

³ Mohammed VI University of Sciences and Health (UM6SS), Casablanca, Morocco

⁴ Emergency Department, Military Hospital Mohammed V, Rabat Medical and Pharmacy School, Mohammed V University, Rabat, Morocco

Abstract

Background: Genomic sequence similarity comparison is a crucial research area in bioinformatics. Multiple Sequence Alignment (MSA) is the basic technique used to identify regions of similarity between sequences, although MSA tools are widely used and highly accurate, they are often limited by computational complexity, and inaccuracies when handling highly divergent sequences, which leads to the development of alignment-free (AF) algorithms.

Results: This paper presents TreeWave, a novel AF approach based on frequency chaos game representation and discrete wavelet transform of sequences for phylogeny inference. We validate our method on various genomic datasets such as complete virus genome sequences, bacteria genome sequences, human mitochondrial genome sequences, and rRNA gene sequences. Compared to classical methods, our tool demonstrates a significant reduction in running time, especially when analyzing large datasets. The resulting phylogenetic trees show that TreeWave has similar classification accuracy to the classical MSA methods based on the normalized Robinson-Foulds distances and Baker's Gamma coefficients.

Conclusions: TreeWave is an open source and user-friendly command line tool for phylogeny reconstruction. It is a faster and more scalable tool that prioritizes computational efficiency while maintaining accuracy. TreeWave is freely available at <https://github.com/nasmaB/TreeWave>.

Keywords: Genome comparison, Alignment-free, Genomic signal processing, Whole-genome phylogeny, DNA embedding, Frequency chaos game representation, K-mer, Discrete wavelet transform

Background

Sequence comparison is fundamental in bioinformatics and genomics, it is used to determine similarities, differences, and evolutionary relationships between DNA or protein sequences. Alignment-based similarity analysis is a key step in genomic sequence comparison, this approach involves comparing sequences according to a scoring system



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

that considers matches, mismatches, and gaps between sequences. Needleman-Wunsch [1] and Smith-Waterman [2] algorithms are among the first dynamic programming algorithms applied to biological sequence alignment. Subsequently, several tools were implemented, whether for sequence similarity searches, such as BLAST [3] and FASTA [4], or multiple sequence alignment (MSA), such as Clustal W [5], MAFFT [6], progressive MAUVE [7], and Muscle [8].

MSA methods are widely used and known to be very accurate, however, they have several limitations [9]; they are time and memory-consuming and can be an NP-hard problem while analysing multiple and large genomes [10, 11]. Furthermore, current alignment-based methods face challenges in identifying the correct homologous positions in highly divergent sequences, this can lead to potential inaccuracies in phylogenetic analysis [9, 12]. Additionally, MSA-based methods struggle to scale with the vast data sets available today; for example, aligning long DNA sequences of millions of nucleotides, such as whole bacterial genomes is practically unfeasible [13]. This has led to the development of alignment-free (AF) methods, especially for comparing genome-scale sequences [14–18]. These methods can be categorized into several groups; the most known are word-based methods and information-theory-based methods [19].

The graphical and numerical representation of genomic sequences is an important process as it is the first step in AF approaches. These representations can be categorized into three types: single-value mapping, multidimensional sequence mapping, and cumulative sequence mapping [20]. Frequency Chaos Game Representation (FCGR) is a DNA encoding method, extended from the Chaos Game Representation (CGR) mapping technique [21]. FCGR maps a one-dimensional sequence into a higher dimensional space based on the k-mers frequencies in the sequence [22]. CGR and FCGR have several applications in bioinformatics, such as phylogenetic analysis, the development of alignment-free approaches, and feature encoding for machine learning [23].

Furthermore, the numerical representation of DNA sequences also enables the application of digital signal processing techniques for analyzing genomic data; which is known as Genomic Signal Processing (GSP) [24]. Recently, the GSP field has attracted researchers' interest, and its techniques are applied in various applications including DNA sequence clustering [25, 26], protein-coding region detection [27], and the development of alignment-free methods [24].

In this paper, we present TreeWave, a user-friendly command line tool for alignment-free analysis based on FCGR transformation and Discrete Wavelet Transform (DWT) of DNA sequences. We have tested our method on different genome types, and the results indicate the proposed method's effectiveness and potential to infer accurate phylogenetic trees.

The first step of the TreeWave approach is mapping DNA sequences to numerical representations. We opted for graphical representations of DNA following the Frequency Chaos Game Representation (FCGR) technique. A noteworthy feature of this technique is that it transforms sequences of different lengths into equal-size images, where each pixel corresponds to the frequency of a particular k-mer in the sequence. Algorithm 1 illustrates the steps to implement FCGR for a given DNA sequence.

Genomic signal processing: discrete wavelet transform

The Discrete Wavelet Transform (DWT) is a mathematical method that breaks down a signal into multiple coefficients. Each group of coefficients represents a level of detail or approximation of the original signal. Using DWT, we can effectively analyze DNA input signals at different resolution levels, capturing both frequency and location information.

Numerical representations of DNA sequences obtained in the first step (Algorithm 1) are considered as digital signal inputs; for each sequence, we applied the Haar Discrete Wavelet Transform up to 5 levels of decomposition (Eq. 1).

$$W(S) = \text{Haar}(fcgr_S, L) \quad (1)$$

where $W(S)$ denotes the wavelet feature vector of the DNA sequence S , $fcgr_S$ is the DNA embedding of the sequence S obtained according to algorithm 1, and L is the decomposition level, which we have set to five.

We opted for a 5 level of decomposition as it provides a good balance between capturing both high and low frequency signal components while maintaining computational efficiency; several studies have also chosen this level of decomposition for its effectiveness [28–31]. At each level of Wavelet decomposition, the FCGR image signal is decomposed into approximation coefficients (ACs) and detail coefficients (DCs); ACs preserve most of the energy from the original signal, capturing its overall characteristics. In contrast, DCs primarily represent specific features of the signal, highlighting detailed variations [32]. The total number of features extracted is the concatenation of all coefficients from the different levels of decomposition. Specifically, in our implementation, Wavelet coefficients are flattened into a single feature vector for each FCGR image. Given a 64×64 FCGR matrix, the number of features is a combination of coefficients across the 5 levels, leading to a highly detailed feature space.

The implementation of DWT is performed by the PyWavelets [33] python module.

Distance matrix computation

In the previous step, we obtained the discrete wavelet feature vector $W(S)$ for each sequence in the input dataset. Subsequently, we constructed the distance matrix of the genomic sequences by computing the pairwise cosine distances between their wavelet feature vectors.

Cosine similarity between two given feature vectors is the cosine of the angle between those two vectors. Hence, considering two DNA sequences $S1$ and $S2$, their cosine distance is defined by Eq. 2, where $\vec{W(S1)}$ and $\vec{W(S2)}$ are the wavelet feature vectors of $S1$ and $S2$ obtained by Eq. 1.



Fig. 1 The workflow of TreeWave

Table 1 Datasets summary

Dataset	Genomes type	Number of sequences	Diversity groups	Average sequences length
Papillomavirus	Human papillomavirus complete genome	146	12 Genotypes: 6 – 11 – 16 – 18 – 31 – 33 – 35 – 45 – 52 – 53 – 58- 66	7926 bp
Hepatitis B	Hepatitis B virus complete genome	87	8 Genotypes: A – B – C – D – E – F – G – H	3200 bp
Streptococcus	Streptococcus bacteria whole genome	31	4 Species: Aglactiae – Pyo- genes – Mutans – Pneumo- niae	2.06 Mb
16 S	16 S ribosomal DNA	13	4 Genera of bacteria: Escheri- chia coli – Streptococcus— Bacillus—Thermus	1518 bp
Mitochondrial DNA	Human mitochondrial complete genome	142	16 Haplogroups: X – U6 – U5 – HV – H1 – R – U3 – K – W – N – T – J – M – L3 – L2 – L1	16,569 bp

$$\text{Cosine_distance}_{S1,S2} = 1 - \frac{\rightarrow W(S1) \cdot \rightarrow W(S2)}{\rightarrow ||W(S1)|| \rightarrow ||W(S2)||} \tag{2}$$

Phylogenetic tree inference

e phylogenetic tree was established using the hierarchical clustering algorithm UPGMA (Unweighted Pair Group Method with Arithmetic Mean); it is an agglomerative clustering approach commonly used to construct dendrograms representing the evolutionary relationships between a set of genomes.

e UPGMA algorithm takes as input the constructed cosine distance matrix and returns the inferred phylogenetic tree in newick format.

A graphical summary of TreeWave workflow is shown in Fig. 1.

Results and discussion

Datasets

Five datasets of di erent sizes and genome types are used in our experimental evaluation, namely, papillomavirus sequences, hepatitis B sequences, streptococcus sequences, 16 S sequences, and mitochondrial DNA sequences. Table 1 contains information on

each dataset, including its size, diversity groups, and average sequence length. The sequences constituting the datasets are publicly available, and the NCBI accession numbers are listed in additional file 1.

DNA sequence to digital signal

Visual representation of DNA sequences enhances the comprehension of genetic information by revealing patterns, similarities, and relationships that might be undetectable through raw sequences, facilitating clustering and classification tasks [34, 35]. Additionally, this representation can help in the application of Machine Learning models by transforming complex sequences into high dimensional features [36–38]. However, visual representations of DNA sequences might obscure critical functional elements such as specific nucleotide positions, sequence motifs and structural components; these techniques may not be effective in study cases requiring detailed functional or positional information [39].

Concerning our proposed method, we opted for an FCGR transformation of DNA sequences, a technique derived from Chaos Game Representation (CGR) which is a 2D graphical representation. A key aspect of CGR is that each point's position encodes the historical information of the preceding DNA sequence, while also visually representing the frequencies of nucleotide patterns. CGR retains the statistical properties of DNA sequences, enabling the exploration of both local and global patterns [40].

In the Frequency Chaos Game Representation (FCGR) of DNA sequences step, each pixel represents a specific k-mer, this means that a k value of 3, for example, indicates that each pixel uniquely represents a subsequence of 3 oligonucleotides, enabling the enumeration of occurrences of all oligonucleotides. For example, Fig. 2 represents the resulting FCGR images of 16 S ribosomal DNA from the following species: *Escherichia Coli*, *Thermus Filiformis*, *Streptococcus Cameli*, and *Bacillus Australimaris* at $k=7$.

Given a multi-fasta file of DNA genomic sequences as input, the first step of our approach is the FCGR transformation of each sequence, this requires setting the right value of k; the k-mer size. This is crucial for accurate analysis because it impacts the resolution and information content of the representation; smaller values provide higher resolution but might lead to sparse data, whereas a larger value of k increases data density but might lose detailed information [41, 42]. Since k-mer length is a crucial parameter in AF phylogenetic inference, researchers have developed standardized approaches for the optimal selection of k values, as KITSUNE software [43].

To determine a range of optimal k values, we have adopted a strategy that depends on the types of genomes to be analyzed, as the optimal k value varies depending on the genome type and size.

We have selected one genome from each dataset (Table 1), and for an interval of k ranging from 1 to 17, we determined the total number of unique k-mers that can be formed (Possible k-mers), and the number of k-mers that appear once in the genomic sequence (Distinct k-mers). According to the results (Fig. 3), we can see that the viruses' genome curves display a similar format; both for the human papillomavirus genome and hepatitis B virus genome, before $k=5$, the number of distinct k-mers is almost 0, and for $k>11$, nearly every possible k-mer is distinct. Therefore, we suggest that the interval 5–11, where there is a progressive growth of possible and distinct k-mers, could contain

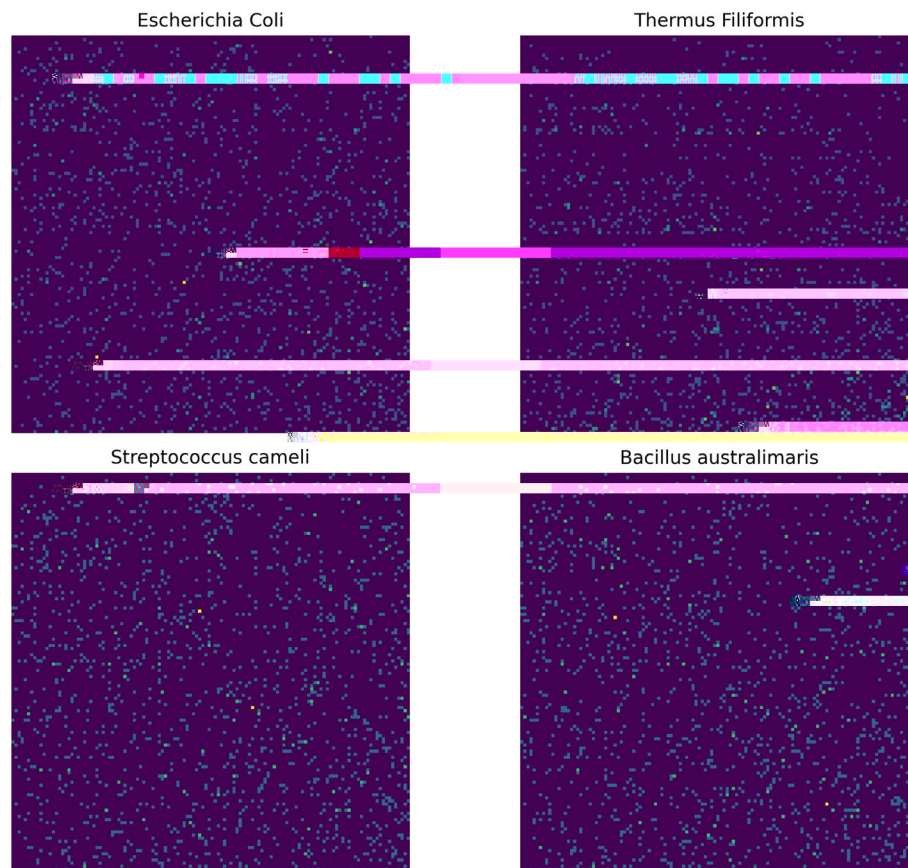


Fig. 2 FCGR images of 16 S ribosomal DNA genomes of four distinct species at $k = 7$

the optimal k value. The range considered for 16 S ribosomal DNA is 4–9, and the range considered for human mitochondrial genomes is 6–11. Whereas for the complete streptococcus genomes, the interval is set from $k = 7$ to $k = 13$.

For each dataset, we have run our workflow at odd k values belonging to the defined intervals and calculated accuracy metrics to specify a k value for the final phylogenetic tree inference of each dataset (Table 2). Additional details of these metrics are provided in the subsequent section about accuracy evaluation.

As mentioned in the implementation section, digital transformations of genomic sequences are obtained according to a specific k value for each dataset, then the cosine distance matrices between wavelet feature vectors are constructed and used as inputs for phylogenetic tree reconstruction.

Phylogenetic tree of human papillomavirus genomes

Human papillomavirus (HPV) is a circular DNA virus from the Papillomaviridae family that causes various epithelial lesions and cancers, predominantly affecting cutaneous and mucosal surfaces [44]. HPV is classified into over 150 different genotypes, which are grouped into five main genera: Alpha, Beta, Gamma, Mu, and Nu, some genotypes are associated with pathologies, hence the importance of studying the phylogeny of this virus [45]. 146 complete human papillomavirus genomes were downloaded from the

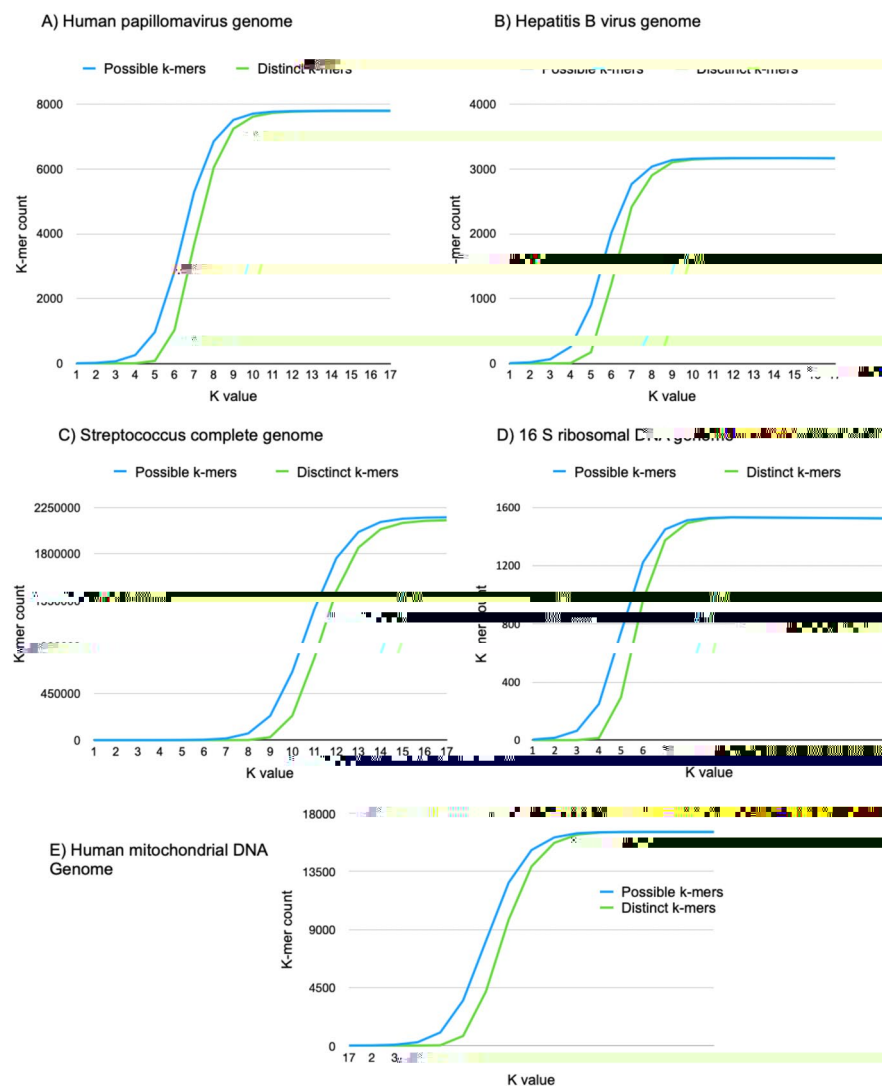


Fig. 3 Distribution of possible and distinct k-mers within k range values from 1 to 17

National Center for Biotechnology Information (NCBI), belonging to 12 different genotypes. The optimal k-mer size is set to 11 in the DNA embedding step. The phylogenetic tree resulting from the proposed method successfully organized HPV genomes into distinct clusters based on their genotypes (Fig. 4A). Figure 4B is the phylogenetic tree of the 146 complete papillomavirus genomes constructed by the alignment-based method Clustal W. Both trees display identical topology, thus reinforcing the validity of Tree-Wave method.

Phylogenetic tree of hepatitis B genomes

The Hepatitis B virus (HBV) genome is a small and enveloped DNA virus that belongs to the Hepadnaviridae family, HBV is classified into 10 main genotypes, designated A through J [46]. The classification of HBV virus genomes provides valuable insights into the impact of specific genotypes on the severity and progression of hepatitis B disease

Table 2 Normalized Robinson Foulds distance and Baker's gamma coefficient results

Dataset	K value	Normalized Robinson Foulds distance	Bakers's Gamma coefficient
Human papillomavirus genomes	5	0.15	0.9728466
	7	0.17	0.9729083
	9	0.15	0.9696713
	11	0.11	0.9954782
Hepatitis B genomes	5	0.40	0.7994529
	7	0.11	0.9997261
	9	0.12	0.9997109
	11	0.10	0.9998684
Streptococcus genomes	7	0.50	0.9437716
	9	0.50	0.9438471
	11	0.43	0.9441479
	13	0.39	0.9874893
16 S ribosomal DNA	5	0.10	0.997625
	7	0.10	0.997625
	9	0.10	0.997625
Human mitochondrial DNA genomes	7	0.27	0.9451289
	9	0.16	0.8946325
	11	0.21	0.8787195

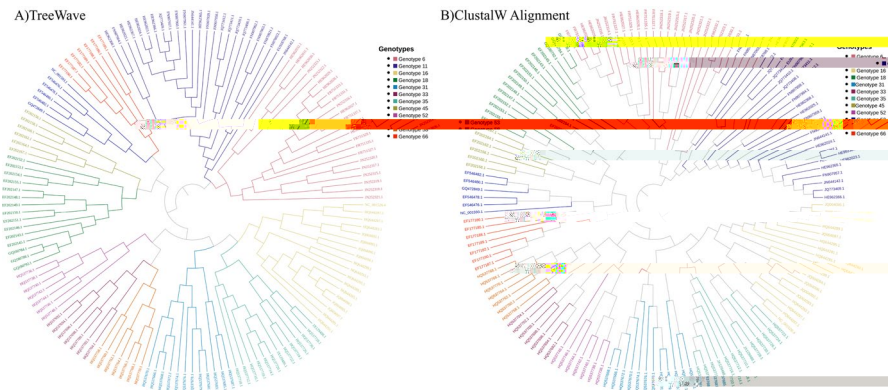


Fig. 4 Phylogenetic tree of 146 whole human papillomavirus complete genome constructed by **A** TreeWave at k = 11 and **B** Clustal W multiple sequence aligner

[47]. HBV dataset used to validate our approach contains 87 complete genomes belonging to 8 distinct genotypes. Figure 5 illustrates the phylogenetic trees of HBV genomes inferred by our proposed method TreeWave, with a k-mer size set to 11 (Fig. 5A) and Clustal W method (Fig. 5B); both trees show accurate grouping of genotypes.

Phylogenetic tree of streptococcus genomes

Classical alignment solutions become inefficient in the analysis of whole-genome bacteria, this is due to the computational intensity and the significant time required for alignment processes, and the difficulty in aligning genomes that are highly similar but have significant differences in gene content and order. We applied our method to 31 complete

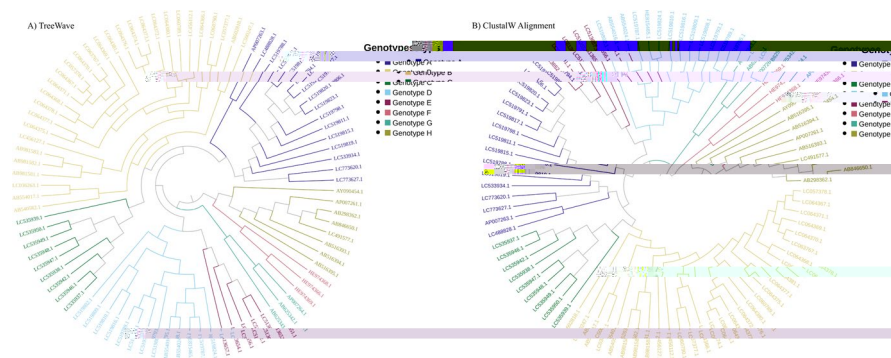


Fig. 5 Phylogenetic tree of 87 hepatitis B virus complete genome constructed by **A** TreeWave at k=11 and **B** Clustal W multiple sequence aligner

streptococcus genomes belonging to 4 different species (Agglactiae, Pyogenes, Mutans, and Pneumoniae) with an average sequence length of 2.06 million bases. The phylogenetic tree generated by our method ($k=13$) is shown in Fig. 6A, we can see that our method accurately classifies the genomes into species. Figure 6B is the phylogenetic tree of the 31 streptococcus genomes generated by the alignment-based tool Mauve. The dendrogram produced by our method aligns closely with the phylogenetic tree derived from the alignment-based method, with the sole discrepancy lying in the topology of interspecies relationships.

Phylogenetic tree of 16 S ribosomal DNA genomes

16 S rRNA gene is an essential marker in bacterial phylogenetics due to its low evolution rate and high conservation across different bacterial species [48]. To test our method,

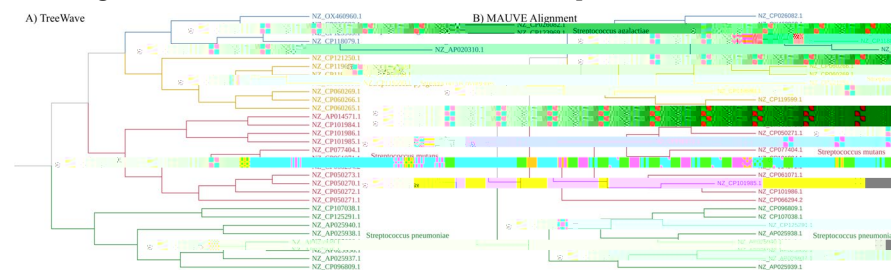


Fig. 6 Phylogenetic tree of 31 streptococcus complete genome constructed by **A** TreeWave at $k = 13$ and **B** Mauve multiple sequence aligner

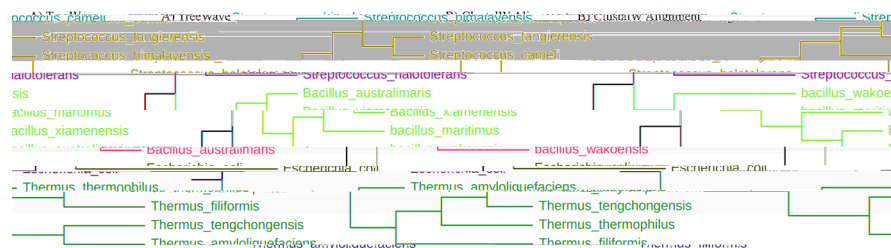


Fig. 7 Phylogenetic tree of 13 16S ribosomal DNA genome constructed by **A** TreeWave at k=9 and **B** Clustal W multiple sequence aligner

we used a dataset of 13 bacterial 16 S ribosomal DNA of 4 distinct groups. In Fig. 7, we present the dendrogram generated by our approach at $k=9$ (Fig. 7A), alongside the phylogenetic tree inferred by Clustal W method (Fig. 7B); we can also see the overall agreement between our proposed tool result and the alignment-based tree. There is only a difference in topology at the level of the thermus clade; the tree generated by the alignment-based method groups thermophilus and filiformis species in one clade, which is not the case in the dendrogram generated by TreeWave.

Phylogenetic tree of human mitochondrial DNA genomes

The human mitochondrial genome is a 16,569 base pair (bp) circular double-stranded DNA molecule. The diversity of human mitogenomes is classified by haplogroups; a set of alphanumeric labels that are implied in various applications such as population genetics, forensics, and studies of disease associations [49]. We applied our method to a dataset containing 142 human mitochondrial genomes, then we identified their haplogroups by Haplogrep2 tool [50]. Figure 8 represents the phylogenetic trees inferred by our method at $k=9$ (Fig. 8A) and Clustal W method (Fig. 8B); both dendrograms accurately classified the genomes according to their haplogroups, with only minor differences in topology observed between the two trees.

Accuracy evaluation and phylogenetic tree distance

To assess the accuracy of our alignment-free approach, we calculate the normalized Robinson Foulds (nRF) distances and Baker's Gamma coefficients between the dendrograms generated by TreeWave and those generated by alignment-based methods. RF distance is a widely used metric for comparing phylogenetic trees, it is the number of splits that differ between two trees [51]. The Baker's Gamma coefficient between two dendrograms quantifies the level of agreement in hierarchical clustering structures [52]. The two metrics are calculated for a range of k values, results are shown in Table 2. Regarding nRF distance, values close to 0 suggest that the trees are very similar in terms of topology, and values close to 1 indicate that the trees are dissimilar; we note that the nRF values obtained don't reach 0.50. About Baker's Gamma coefficient, all values are close to 1, which indicates that dendrograms generated by TreeWave and those generated by classical alignment-based methods have a perfect match in terms of clustering structure.

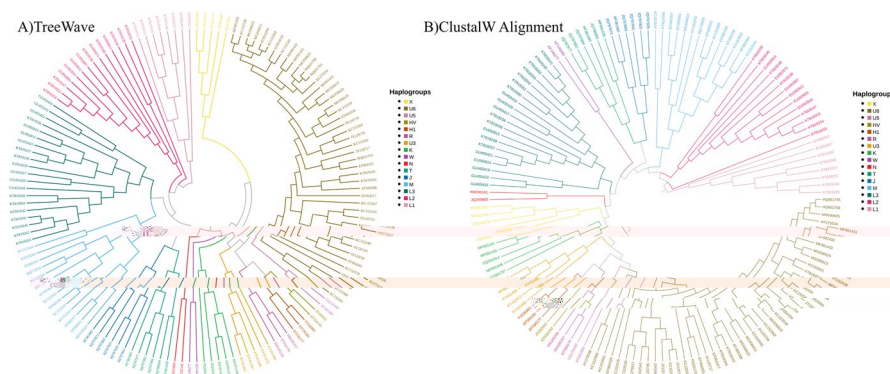


Fig. 8 Phylogenetic tree of 142 whole human mitochondrial DNA genome constructed by **A** TreeWave at $k=9$ and **B** Clustal W multiple sequence aligner

Time performance

To assess the computational efficiency of TreeWave, we evaluated the run time cost of the analyzed datasets using both the classical alignment methods and our proposed alignment-free approach TreeWave (Table 3). The results demonstrate that TreeWave has a significant advantage in terms of time efficiency, especially when analyzing large genomic datasets; it achieves speedup factors of 15.6 and 1160.37 against Clustal W alignment method for the human papillomavirus dataset and human mitochondrial DNA dataset respectively. Lower speed factors were observed when comparing TreeWave performance against MAFFT; TreeWave proved to be significantly faster by factors of 5.89 and 436.59 for the papillomavirus and mtDNA datasets, respectively. For the hepatitis B dataset, TreeWave is approximately two times faster than Clustal W and one time faster than MAFFT. In the case of a smaller dataset, such as 16 S ribosomal DNA, the three tools performed almost similarly. Another pronounced difference was observed with the whole genome streptococcus bacteria dataset, TreeWave completed the process within a reasonable timeframe of 48 min, while multiple sequence alignment remains unfinished for this dataset by Clustal W and MAFFT, which led us to use Progressive Mauve alignment that took more than 19 h for the execution time. These analyses were performed on a MacBook with an Apple M1 chip and 8GB of memory.

Performance comparison with alignment-free methods

To assess the performance of TreeWave, we compared its results with several state-of-the-art Alignment-Free tools including Filtered Spaced Word Matches (FSWM) [15], *k*-mer inner distance distribution for phylogenetic analysis (KINN) [53], Alignment-free Dissimilarity Analysis & Comparison Tool (ADACT) [54], and an Alignment-Free Phylogeny Estimation Method Using Cosine Distance on Minimal Absent Word Sets (CD-MAWS) [55].

Table 3 Run time benchmark

Dataset	No of sequences	Method	Run time
Human papillomavirus	146	TreeWave	13.74 min
		Clustal W	3.59 h
		MAFFT	1.3473 h
Hepatitis B virus	87	TreeWave	3.5505 min
		Clustal W	10.51 min
		MAFFT	4.9542 min
Streptococcus bacteria	31	TreeWave	48 min
		Clustal W	–
		MAFFT	–
		Mauve	19.68 h
16 S ribosomal DNA	13	TreeWave	10.94 s
		Clustal W	21.11 s
		MAFFT	19.18 s
Human mtDNA	142	TreeWave	39.06 s
		Clustal W	12.59 h
		MAFFT	5.03 h

We applied these tools to construct phylogenetic trees of the five datasets separately (Table 1). FSWM, ADACT and CD-MAWS produce a phylogenetic tree in newick format, whereas KINN result is a pairwise distances matrix, which we imported into MEGA software [56] and performed UPGMA analysis to obtain the tree. We then calculated the normalized Robinson-Foulds distance between each resulting tree and its reference tree; the results are represented in Table 4. According to nRF values, we note that TreeWave consistently demonstrates a good performance across the five datasets when compared to other methods. TreeWave achieved the best performance for Hepatitis B genomes, Streptococcus genomes, 16S ribosomal DNA, and human mitochondrial DNA genomes. On the human papillomavirus genomes dataset, TreeWave performed well with an nRF value of 0.15; the third lowest value after ADACT and CD-MAWS. We were unable to obtain phylogenetic trees for the complete Streptococcus genomes using ADACT, as the web server provided for this tool imposes a sequence length limit. Similarly, we could not generate results with KINN, likely since this tool was not tested on complete bacterial genomes. Overall, TreeWave showed competitive performance across diverse datasets, often outperforming other state-of-the-art tools, and showed comparable results on specific datasets with CD-MAWS.

Numerous Alignment-Free approaches for sequence comparison have been developed, these approaches include methods based on Markov chain model to estimate the relationships between DNA sequences [57], graph theory and nucleotide triplets [58], k-mer forest structures of DNA sequences [59], and triplet frequencies [60]. However, a limitation of many alignment-free methods is that, while authors explain and validate their approaches, they often don't implement a publicly available tool for testing. To advance this field, researchers should be encouraged to produce accessible tools, and open-source development is particularly important for fostering further innovation and collaboration. In response to these limitations, recent efforts have focused on benchmarking studies of proposed alignment-free methods to assess their effectiveness and robustness [9, 61].

To further validate our approach, we used an additional dataset of 25 complete mitochondrial DNA sequences of fish samples, this is a benchmarking dataset provided by Afproject [61]; it's a publicly available framework that developers of AF methods could use to evaluate their approaches. We uploaded the pairwise distance matrix generated by treeWave at $k = 9$ to Afproject server for evaluation, then according to the benchmark report generated by Afproject, among 107 methods with 18

Table 4 nRF distance comparison between phylogenetic trees constructed using 5 alignment-free methods and reference trees

Dataset	TreeWave	FSWM	KINN	ADACT	CD-MAWS
Human papillomavirus	0.15	0.87	0.23	0.11	0.09
Hepatitis B virus	0.10	0.35	0.11	0.14	0.11
Streptococcus bacteria	0.39	0.89	–	–	0.41
16 S ribosomal DNA	0.10	0.20	0.12	0.10	0.10
Human mtDNA	0.16	0.97	0.47	0.28	0.26

possible ranks due to ties in accuracy, Treewave is ranked 2th, with a nRF value of 0.09 and a normalized Quartet Distance (nQD) value of 0.0327.

Conclusions

This paper presents TreeWave, an alignment-free approach for phylogenetic tree inference of DNA genome datasets. The method is based on Frequency Chaos Game Representation of DNA sequences and Discrete Wavelet Transform as signal processing technique. TreeWave is tested on different datasets; the obtained dendrograms accurately classify the genomes into their diversity groups. The effectiveness of TreeWave is also proved by comparing it with alignment-based methods and state-of-the-art Alignment-Free methods; the normalized Robinson-Foulds distances obtained underscore the ability of TreeWave to accurately capture evolutionary relationships among sequences. In terms of time performance, TreeWave approach outperformed alignment-based methods across diverse datasets, exhibiting faster execution times. Beyond its primary functionality of inferring phylogenetic trees, TreeWave stands out for its open-source nature, allowing researchers to tailor it to specific needs. For example, users can employ the FCGR transformation algorithm to generate images suitable for machine-learning analyses or for visualizing genome structures. Furthermore, the pairwise distance matrix computation feature can be used for genome clustering or genetic diversity analysis.

We aim for upcoming TreeWave releases to incorporate a web server to enhance user-friendliness, and simplify the process of selecting the optimal K value.

Abbreviations

MSA	Multiple Sequence Alignment
AF	Alignment Free
FCGR	Frequency Chaos Game Representation
DWT	Discrete Wavelet Transform
nRF	Normalized Robinson-Foulds
CGR	Chaos Game Representation
GSP	Genomic Signal Processing
DWT	Discrete Wavelet Transform
AC	Approximation Coefficient
DC	Detail Coefficient
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
HPV	Human papillomavirus
NCBI	National Center for Biotechnology Information
HBV	Hepatitis B virus
nQD	Normalized Quartet Distance

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-024-05992-3>.

Supplementary material 1

Acknowledgements

Not applicable.

Author contributions

NB: Conception and design of TreeWave, software implementation, software testing, manuscript drafting, visualization. HB: Software testing, manuscript drafting, visualization. LB: Software testing, manuscript drafting. AI: Project supervision, conception and design of TreeWave, software testing, manuscript drafting. All authors read and approved the final manuscript.

Funding

This work was carried out under National Funding from the Moroccan Ministry of Higher Education and Scientific Research to AI (PPR1). This work was also supported, by a grant to AI from the Institute of Cancer Research of the foundation Lalla Salma, and also by a grant from Biocodex Microbiota Foundation.

Availability of data and materials

Both source code and datasets are available in the GitHub repository <https://github.com/nasmaB/TreeWave>, and at Zenodo repository (<https://doi.org/10.5281/zenodo.13739906>). The sequences constituting the datasets are publicly available, and the NCBI accession numbers are listed in additional file 1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 10 September 2024 Accepted: 18 November 2024

Published online: 27 November 2024

References

- Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 1970;48(3):443–53.
- Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol.* 1981;147(1):195–7.
- Altschul SF, Madden TL, Schäfer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–402.
- Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci.* 1988;85(8):2444–8.
- Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994;22(22):4673–80.
- Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 2002;30(14):3059–66.
- Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 2004;14(7):1394–403.
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792–7.
- Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* 2017;18(1):186.
- Just W. Computational complexity of multiple sequence alignment with SP-score. *J Comput Biol.* 2001;8(6):615–23.
- Wang L, Jiang T. On the complexity of multiple sequence alignment. *J Comput Biol.* 1994;1(4):337–48.
- Ranwez V, Chantret NN. Strengths and Limits of Multiple Sequence Alignment and Filtering Methods. Chapter 2.2.2:1–2.2:36.
- Bernard G, Chan CX, Chan YB, Chua XY, Cong Y, Hogan JM, et al. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Brief Bioinform.* 2017;20(2):426–35.
- Comin M, Verzotto D. Alignment-free phylogeny of whole genomes using underlying subwords. *Algorithms Mol Biol.* 2012;6(7):34.
- Leimeister CA, Sohrabi-Jahromi S, Morgenstern B. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics.* 2017;33(7):971–9.
- Yang Young Lu, Tang K, Ren J, Fuhrman JA, Waterman MS, Sun F. CAFE: accelerated alignment-free sequence analysis. *Nucleic Acids Res.* 2017;45(W1):W554–9. <https://doi.org/10.1093/nar/gkx351>.
- Zuo G, Hao B. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. *Genom Proteomics Bioinform.* 2015;13(5):321–31.
- Jun SR, Sims GE, Wu GA, Kim SH. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution. *Proc Natl Acad Sci.* 2010;107(1):133–8.
- Ren J, Bai X, Lu YY, Tang K, Wang Y, Reinert G, et al. Alignment-free sequence analysis and applications. *Annu Rev Biomed Data Sci.* 2018;1:93–114.
- Mendizabal-Ruiz G, Román-Godínez I, Torres-Ramos S, Salido-Ruiz RA, Morales JA. On DNA numerical representations for genomic similarity computation. *PLoS ONE.* 2017;12(3): e0173288. <https://doi.org/10.1371/journal.pone.0173288>.
- Je rey HJ. Chaos game representation of gene structure. *Nucleic Acids Res.* 1990;18(8):2163–70.
- Wang Y, Hill K, Singh S, Kari L. The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene.* 2005;14(346):173–85.
- Löchel HF, Heider D. Chaos game representation and its applications in bioinformatics. *Comput Struct Biotechnol J.* 2021;10(19):6263–71.

24. Borrayo E, Mendizabal-Ruiz EG, Vélez-Pérez H, Romo-Vázquez R, Mendizabal AP, Morales JA. Genomic signal processing methods for computation of alignment-free distances from DNA sequences. *PLoS ONE*. 2014;9(11): e110954.
25. Mendizabal-Ruiz G, Román-Godínez I, Torres-Ramos S, Salido-Ruiz RA, Vélez-Pérez H, Morales JA. Genomic signal processing for DNA sequence clustering. *PeerJ*. 2018;24(6): e4264.
26. Bao JP, Yuan RY. A wavelet-based feature vector model for DNA clustering. *Genet Mol Res*. 2015;14(4):19163–72.
27. Mabrouk MS. Advanced genomic signal processing methods in DNA mapping schemes for gene prediction using digital filters. *Am J Signal Process*. 2017;7(1):12–24.
28. Daud SNSS, Sudirman R. Decomposition Level Comparison of Stationary Wavelet Transform Filter for Visual Task Electroencephalogram | Jurnal Teknologi (Sciences & Engineering). 2015 May 28. Available from: <https://journals.utm.my/index.php/jurnalteknologi/article/view/4661>
29. S. Chopra, H. Kaur and A. Kaur. 2010 Selection of best wavelet basis for image compression at decomposition level 5. 2010 2nd international conference on computer technology and development, Cairo, Egypt, pp. 442–445. <https://doi.org/10.1109/ICCTD.2010.5645837>
30. Srivastava V, Purwar RK. A five-level wavelet decomposition and dimensional reduction approach for feature extraction and classification of MR and CT scan images. *Appl Computational Intell Soft Comput*. 2017;9(1):9571262.
31. Saini S, Dewan L. Application of discrete wavelet transform for analysis of genomic sequences of *Mycobacterium tuberculosis*. *Springerplus*. 2016;5:64. <https://doi.org/10.1186/s40064-016-1668-9>.
32. H. K. Kwan and S. B. Arniker. Numerical representation of DNA sequences. 2009 *IEEE International conference on electro/information technology*, windsor, ON, Canada, 2009, pp. 307–310. <https://doi.org/10.1109/EIT.2009.5189632>.
33. Lee GR, Gommers R, Waselewski F, Wohlfahrt K, O'Leary A. PyWavelets: a python package for wavelet analysis. *J Open Source Softw*. 2019;4(36):1237.
34. Bashir M, Mathur R. Graphical Representation of a DNA Sequence and Its Applications to Similarities Calculation: A Mathematical Model. In: Sahni M, Merigó JM, Jha BK, Verma R, editors. *Mathematical Modeling Computational Intelligence Techniques and Renewable Energy Advances in Intelligent Systems and Computing*. Singapore: Springer; 2021.
35. S. N. Hossain, M. H. Kabir and A. Pal, "Alignment Free Sequence Similarity Estimation using Local Binary Pattern on DNA Trajectory Images," 2021 *Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, Kitakyushu, Japan, 2021, pp. 1–7. <https://doi.org/10.1109/ICIEV-icIVPR52578.2021.9564141>.
36. LA Santamaría C, Zuñiga HS, Pineda TIH, Somodevilla MJ, Rossainz LM. DNA sequence recognition using image representation. *RCS*. 2019;148(3):105–14.
37. Yin B, Balvert M, Zambrano D, Schönhuth A, Bohte S. An image representation based convolutional network for DNA classification [Internet]. arXiv: 2018. Available from: <http://arxiv.org/abs/1806.04931>
38. Löchel HF, Eger D, Sperlea T, Heider D. Deep learning on chaos game representation for proteins. *Bioinformatics*. 2020;36(1):272–9.
39. Jin X, Jiang Q, Chen Y, Lee SJ, Nie R, Yao S, et al. Similarity/dissimilarity calculation methods of DNA sequences: a survey. *J Mol Graph Model*. 2017;1(76):342–55.
40. Yin C. Encoding and decoding DNA sequences by integer chaos game representation. *J Comput Biol*. 2019;26(2):143–51.
41. Swain MT, Vickers M. Interpreting alignment-free sequence comparison: what makes a score a good score? *NAR Genom Bioinform*. 2022;4(3):lqac062.
42. Zhang Q, Jun SR, Leuze M, Ussery D, Nookaew I. Viral phylogenomics using an alignment-free method: a three-step approach to determine optimal length of k-mer. *Sci Rep*. 2017;7(1):40712.
43. Pornputtapong N, Acheampong DA, Patumcharoenpol P, Jenjaroenpun P, Wongsurawat T, Jun SR, et al. KITSUNE: a tool for identifying empirically optimal K-mer length for alignment-free phylogenomic analysis. *Front Bioeng Biotechnol*. 2020;23(8): 556413.
44. Luria L, Cardoza-Favarato G. Human Papillomavirus. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2024. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK448132/>
45. Jendoubi-Ferchichi M, Satouri L, Ghoul F, Malek-Mellouli M, Derbel AM, Makni MK, et al. Phylogeny and classification of human papillomavirus (HPV)16 and HPV18 variants based on E6 and L1 genes in tunisian women with cervical lesions. *Asian Pac J Cancer Prev*. 2018;19(12):3361–6.
46. Cremer J, van Heiningen F, Veldhuijzen I, Benschop K. Characterization of hepatitis B virus based complete genome analysis improves molecular surveillance and enables identification of a recombinant C/D strain in the Netherlands. *Heliyon*. 2023;9(11): e22358.
47. Lin CL, Kao JH, Chen BF, Chen PJ, Lai MY, Chen DS. Application of hepatitis B virus genotyping and phylogenetic analysis in intrafamilial transmission of hepatitis B virus. *Clin Infect Dis*. 2005;41(11):1576–81.
48. Hassler HB, Probert B, Moore C, Lawson E, Jackson RW, Russell BT, et al. Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies. *Microbiome*. 2022;10(1):104.
49. Habbane M, Montoya J, Rhouda T, Sbaoui Y, Radallah D, Emperador S. Human mitochondrial DNA: particularities and diseases. *Biomedicine*. 2021;9(10):1364.
50. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res*. 2016;44:W58–63.
51. Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math Biosci*. 1981;53(1):131–47.
52. Baker FB. Stability of two hierarchical grouping techniques case 1: sensitivity to data errors. *J Am Stat Assoc*. 1974;69(346):440–5.
53. Tang R, Yu Z, Li J. KINN: an alignment-free accurate phylogeny reconstruction method based on inner distance distributions of k-mer pairs in biological sequences. *Mol Phylogenet Evol*. 2023;1(179): 107662.
54. Akon M, Akon M, Kabir M. M Saifur Rahman, M Sohel Rahman, ADAC: a tool for analysing (dis)similarity among nucleotide and protein sequences using minimal and relative absent words. *Bioinformatics*. 2021;37(10):1468–70. <https://doi.org/10.1093/bioinformatics/btaa853>.

55. Anjum N, Nabil RL, Rafi RI, Bayzid MDS, Rahman MS. CD-Maws: an alignment-free phylogeny estimation method using cosine distance on minimal absent word sets. *IEEE/ACM Trans Computational Biol Bioinform.* 2023;20(1):196–205.
56. Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. *Mol Biol Evol.* 2021;38(7):3022–7. <https://doi.org/10.1093/molbev/msab120>.
57. Saw AK, Raj G, Das M, Talukdar NC, Tripathy BC, Nandi S. Alignment-free method for DNA sequence clustering using Fuzzy integral similarity. *Sci Rep.* 2019;9(1):3753.
58. Das S, Das A, Bhattacharya DK, Tibarewala DN. A new graph-theoretic approach to determine the similarity of genome sequences based on nucleotide triplets. *Genomics.* 2020;112(6):4701–14.
59. G. Gamage, N. Gimhana, A. Wickramarachchi, V. Mallawaarachchi and I. Perera. 2019, Alignment-free Whole Genome Comparison Using k-mer Forests. *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, pp. 1–7, <https://doi.org/10.1109/ICTer48817.2019.9023714>.
60. Kirichenko AD, Poroshina AA, Sherbakov DY, Sadovsky MG, Krutovsky KV. Comparative analysis of alignment-free genome clustering and whole genome alignment-based phylogenomic relationship of coronaviruses. *PLoS ONE.* 2022;17(3): e0264640. <https://doi.org/10.1371/journal.pone.0264640>.
61. Zieleszinski A, Girgis HZ, Bernard G, et al. Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* 2019;20:144. <https://doi.org/10.1186/s13059-019-1755-7>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.