Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2022.00065

Research Article

# A new feature vector model for alignment-free DNA sequence similarity analysis

**Emre DELIBAŞ[1,*], Ahmet ARSLAN[2]**

[1]*Department of Computer Engineering, Sivas Cumhuriyet University, Sivas, Türkiye*
[2]*Department of Computer Engineering, Selçuk University, Konya, Türkiye*

## ARTICLE INFO

*Article history*
Received: 08 February 2021
Accepted: 23 March 2021

**Keywords:**
Alignment-free Comparison; DNA Sequence Similarity; Feature Extraction

## ABSTRACT

Improvements in technology have triggered the production of big data. Within this scope, enormous amounts of biological data have been generated. A number of analysis methods have been developed to access the information contained in biological data. DNA sequence analysis has drawn particular attention in recent years. As an alternative to alignment-based sequence comparison methods that have high computational costs, alignment-free comparison methods have emerged. These methods can calculate sequence similarity by applying different dimensions of numerical characterizations. In this paper, we propose a novel alignment-free DNA sequence analysis method based on a feature extraction strategy. The method utilizes numerical characterization and is implemented by calculating mean distance of the transitions, mean distance of the nucleotide duplications, and the base frequencies. The method then measures the similarity between 7-dimensional vectors that are obtained through feature extraction. Using this approach, we conducted a sequence similarity analysis of two different DNA sequence datasets of different lengths to demonstrate the effectiveness of the method. The proposed method shows that a simple and successful feature vector can be obtained when DNA sequences having many properties are used in combination with appropriate and effective descriptors. With this strategy, reasonable results were obtained with a low computational cost.

**Cite this article as:** Emre D, Ahmet A. A new feature vector model for alignment-free DNA sequence similarity analysis. Sigma J Eng Nat Sci 2022;40(3):610–619.

## INTRODUCTION

After the publication of the Human Genome Project [1] , an enormous amount of biological data has emerged as a result of rapid technological development. Most research on the related analytical methods has aimed to extract information contained in this massive amount of data. DNA sequence analysis has the potential to play

***Corresponding author.**
*E-mail address: edelibas@cumhuriyet.edu.tr

*This paper was recommended for publication in revised form by Regional Editor Pravin Katare*

Published by Yıldız Technical University Press, İstanbul, Turkey
Copyright 2021, Yıldız Technical University. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

a key role in these studies. DNA sequence analysis is the first step in identifying similar nucleotide sequences in large genomic data repositories, indicating evolutionary or affinity relationships and/or related pathophysiological processes. DNA sequence similarity analysis is an inference process that compares unknown sequences with known ones to identify the functions of the unknown sequences [2]. Computational methods for DNA sequence similarity analysis offer the most successful solutions for this difficult biological analysis.

In recent years, a number of methods have been proposed to analyze similarities among DNA sequences to correctly identify genetic information [3]. The methods used in DNA sequence similarity analysis can be divided into two basic groups: alignment-based methods and alignment-free methods. There are two groups of alignment-based methods, dynamic programming [4]-based methods [5, 6] and heuristic-based methods [7, 8]. Alignment-based methods use thorough searching, gap insertions, and base shifting in sequences to obtain optimal alignments, but they carry a high computational cost.

The methods mentioned above are different from alignment-free methods, which convert each DNA sequence into a numerical feature vector to quickly compute similarity. By contrast, alignment-free methods are divided into several subcategories based on the applied strategy. Graphical representation-based methods are widely used to analyze and visualize DNA sequences. The first significant example of this method was initiated in 1983 by Hamori and Ruskin [9] who presented a three-dimensional (3D) geometric representation of DNA sequences. Nandy and Randic further developed this method in 1994 and 2003 [10, 11]. As a result of such developments, graphical representation-based methods have found a wide range of 2D [12-15], 3D [16-20], and other applications [21-24]. In addition to graphical representations, several other methods have been used in numerical representations of DNA sequences. Some of these methods use a directed graph to structure the relations between dinucleotides [25, 26]; they can also use an undirected graph to structure a complex network of different nucleotide lengths [27-31]. Methods using word-based measurement are also among the most widely used alignment-free methods [31]. Other commonly used methods convert DNA sequences to numerical vectors by considering the frequency or positional information of nucleotides and by grouping the nucleotides according to their chemical properties. Previous studies \ have also introduced methods that are based on the frequency of nucleotides. Other proposed methods group nucleotides according to their chemical properties[32-35]. In contrast to these studies involving long feature vectors and complex calculations, in which all physicochemical properties are included, in the current study, we propose a novel and simple method based on mean distance of the transitions, mean distance of the nucleotide duplications, and the base frequencies. The main advantage of this method is that it produces a small and easy to compute 7-dimensional feature vector and scans the sequences only once to obtain the numerical values that make up this vector.

## MATERIALS AND METHODS

DNA sequences consist of simple molecules called nucleotides. Nucleotides are the phosphate esters of nucleosides and are the components of DNA. The three components that construct all nucleotides are a nitrogen heterocyclic base, a pentose sugar, and a phosphate residue. The major bases are monocyclic **pyrimidines** (Y) and bicyclic **purines** (R). The major purines are **adenine** (A) and **guanine** (G), and the major pyrimidines are **cytosine** (C), **thymine** (T) [36]. These nucleotides can also be expressed as a text string consisting of the above four letters (A, G, C, and T), which represents the DNA sequence.

Mutation is the change of nucleotide sequence of an organism, virus, extrachromosomal DNA. It can occur on DNA or RNA sequence. One of the important mutations is the point mutation which is the alteration of just one nucleotide. Based on the type of base altered, a point mutation can be classified as a transition or a transversion mutation. A pyrimidine replaced by another pyrimidine (C to T or T to C), or a purine replaced by another purine (A to G or G to A) is a transition mutation. A pyrimidine replaced by a purine, or a purine replaced by a pyrimidine is a transversion mutation. Transition mutations are far more prevalent than transversion mutations [37].

From the perspective of computer science, a DNA sequence is a string statement. Therefore, text analysis methods can be adapted to identify similarities between DNA sequences. When strings of text are compared to measure similarities among them, they are first converted to a vector in different formats, regardless of their length. Finally, similarities between the obtained vectors can be calculated. When creating a feature vector, it is important that the values in the vector should be distinctive for numerical characterization. The distinctiveness of these values will influence the success of the similarity analysis. In this study, we generated a 7-dimensional feature vector for each analyzed DNA sequence. The feature vector uses mean distance of the transitions, mean distance of the nucleotide duplications, and the base frequencies. These different parameters come together to have a powerful effect in characterizing DNA with a small vector.

Information on transitions in the nucleotide bases were used for the first two values of the feature vector. Purines and pyrimidines are symbolized as R and Y, respectively. Thus, a transition will be expressed as RR or YY. Let $\boldsymbol{S} = \boldsymbol{s_1 s_2 s_3 \ldots s_n}$ be a DNA sequence with length n. We obtain a map from this sequence, named $\boldsymbol{\phi_{RY}}$:

$$\phi_{RY}(S) = \phi_{RY}(S_1)\phi_{RY}(S_1)\phi_{RY}(S_1)\ldots\phi_{RY}(S_n),$$

where $\phi_{RY}(S_i) = \begin{cases} R, \ if \ s_i \in Purines \\ Y, if \ if \ s_i \in Pyrimidines \end{cases}, \ i = 1,2...,n$

By applying this notation, we can exemplify the corresponding sequence (*S*) = *AAGCTTATAGGCCCT* as follows:

$$\phi_{RY}(S) = RRRYYYRYRRRYYYY$$

In the proposed method, we handle the transition points by shifting a 2-length window on $\phi_{RY}$. The first two values of the feature vector ($\mu_{RR}$,$\mu_{YY}$) describes mean distances of the transitions from the beginning of $\phi_{RY}$.

To calculate the third value of the feature vector, in the same way, we handle the nucleotide duplication points by shifting a 2-length window on *S*. $\mu_D$ is mean distance of the nucleotide duplications from the beginning of the sequence *S*.

The mean distances $\mu_{RR}$,$\mu_{YY}$, and $\mu_D$ are defined as:

$$\mu_{RR} = \sum_{i=1}^{n_{RR}} \frac{d_i}{n_{RR}} \quad (1)$$

where $d_i$ is the distance from the beginning of $\phi_{RY}$ to *i*th transition occurring as RR, $n_{RR}$ is the number of the transitions occurring as RR.

$$\mu_{YY} = \sum_{i=1}^{n_{YY}} \frac{d_i}{n_{YY}} \quad (2)$$

where $d_i$ is the distance from the beginning of $\phi_{RY}$ to *i*th transition occurring as YY, $n_{YY}$ is the number of the transitions occurring as YY.

$$\mu_D = \sum_{i=1}^{n_D} \frac{d_i}{n_D} \quad (3)$$

where $d_i$ is the distance from the beginning of *S* to *i*th duplication, $n_D$ is the number of the nucleotide duplications in *S*.

If two DNA sequences are similar, the mean distances should also be similar. However, since the present values may be inadequate for comparing DNA sequences and the mean distances of the transitions and the nucleotide duplications in the different positions may be the same, the vector requires further strengthening.

The last four values of the feature vector represent the contents of the nucleotide bases A, G, C, and T in a given DNA sequence. The four numerical parameters indicate the total number of A, G, C and T nucleotides and are symbolized as $n_A$, $n_G$, $n_C$, and $n_T$ respectively. These four integer values are simple, but they are important parameters for sequence characterization.

A combined feature vector that contains different numerical parameters is thus obtained; this vector can be used to analyze the similarity between DNA sequences. The feature vector contains 7-dimensional information and is presented as:

$$V = [\mu_{RR},\mu_{YY}.\mu_D,n_A,n_G,n_C,n_T] \quad (4)$$

For the given DNA sequence *S* and the given $\phi_{RY}(S)$ values, the parameters and the 7-dimensional vector *V* are as follows:

$$\mu_{RR} = \frac{22}{4} = 5,5$$
$$\mu_{YY} = \frac{48}{5} = 9,6$$
$$\mu_D = \frac{41}{5} = 8,2$$
$$n_A = 4, n_G = 3, n_C = 4, n_T = 4$$
$$V = [5.5,9.6,8.2,4,3,4,4]$$

**Similarity Calculation**

In the previous section, we obtained a feature vector V in the 7-dimensional linear space. Obtaining these vectors allows DNA sequences to be compared. In calculating the similarity between DNA sequences, the distance calculation is the basis of the analysis and is an important step to obtain accurate results. Euclidean distances are very often used when making comparisons [3]. The similarity between the previously calculated characterization vectors can be obtained by applying the Euclidean distances between their end points:

$$E = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (5)$$

**Construction of phylogenetic trees**

We performed similarity measurements with "MATLAB Statistics and Machine Learning Toolbox" and the "MATLAB Bioinformatics Toolbox" for the proposed method. We used "pdist" and "seqlinkage" functions to generate phylogenetic trees to analyze the feature vectors we used in clustering. We used "Euclidean distance (default)" for pdist and "single" for seqlinkage as the parameters of the functions. We then generated dendrograms in MATLAB R2018b. MEGA7 (Molecular Evolutionary Genetics Analysis software) was used as a reference to compare to the phylogenetic trees [38].

## RESULTS

**Data Description**

We applied the proposed method to two DNA sequence datasets and compared the results with the reference trees generated by MEGA7. We chose datasets containing

different sequence lengths. The two data sets we used to evaluate our method were used by previous researchers.

## METHOD IMPLEMENTATION

**NADH dehydrogenase subunit 4 genes**

We used the NADH dehydrogenase subunit 4 gene from twelve species of four different groups of primates. All the DNA sequences were obtained from the NCBI genetic database and are given in Table 1. The sequence lengths between 893 to 896 base pairs. They were previously investigated and reported by Hayasaka et al. [16] and subsequently used by Zhang [32, 39], Qi et al. [25], Chen et al. [37], and Delibas et al. [40].

We applied the proposed method to the sequences given in Table 1 and obtained the feature vectors. We then calculated the similarities between these vectors using Euclidean distances. The phylogenetic tree obtained by this calculationis given in Figure 1. We analyzed the same DNA sequences in MEGA7 software by the alignment-based method ClustalW. We used the UPGMA method to construct the phylogenetic tree presented in Figure 2.

The trees generated by the proposed method and the reference method (MEGA7) showed overall qualitative agreement in the similarity matrix. To clarify and visualize this result, in addition, in the distance measurement plots in Figure 3, we revealed the similarity between the human and other species from the sequences given in Table 1. The compatibility of these two curves shows the general agreement between the proposed method and the results obtained with the reference method. In Figure 4, it shows the projection of twelve feature vectors to the 2D property

**Table 1**. NADH dehydrogenase subunit 4 genes of 12 species genome information from NCBI

| | Species | Accession Code | Length (bp) |
|---|---|---|---|
| **1** | *Macaca fascicularis* | M22653 | 896 |
| **2** | *Macaca fuscata* | M22651 | 896 |
| **3** | *Macaca mulatta* | M22650 | 896 |
| **4** | *Macaca sylvanus* | M22654 | 896 |
| **5** | *Saimiri sciureus* | M22655 | 893 |
| **6** | *Chimpanzee* | V00672 | 896 |
| **7** | *Lemur catta* | M22657 | 895 |
| **8** | *Gorilla* | V00658 | 896 |
| **9** | *Hylobates* | V00659 | 896 |
| **10** | *Sumatran Orangutan* | V00675 | 895 |
| **11** | *Tarsius syrichta* | M22656 | 895 |
| **12** | *Human* | L00016 | 896 |

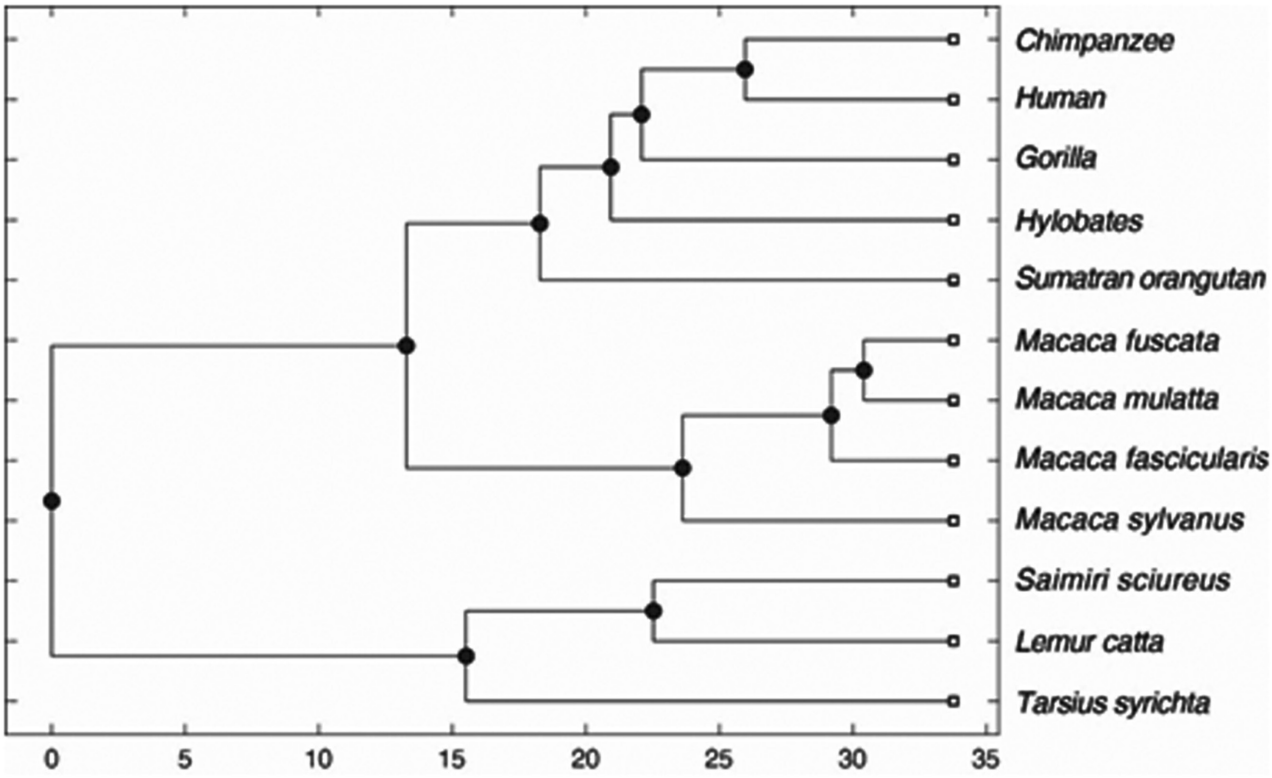


**Figure 1.** Phylogenetic tree generated by the proposed method.

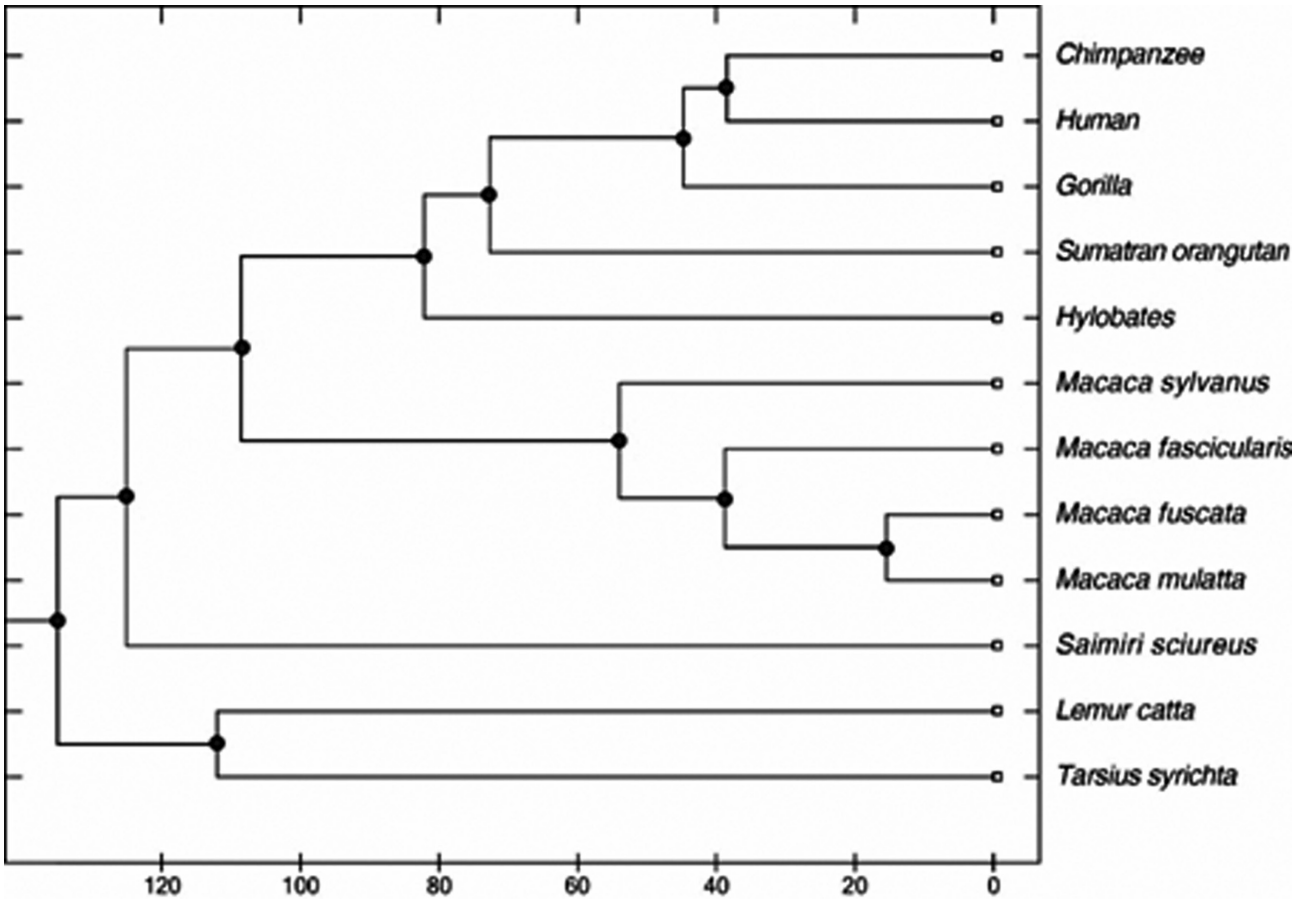


**Figure 2.** Phylogenetic tree generated by MEGA7 based on ClustalW alignment and the UPGMA method.

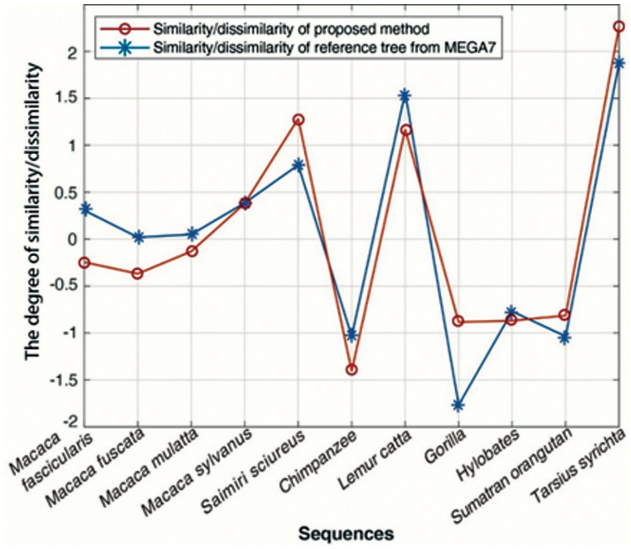


**Figure 3.** The degree of similarity between human and the other 11 species.

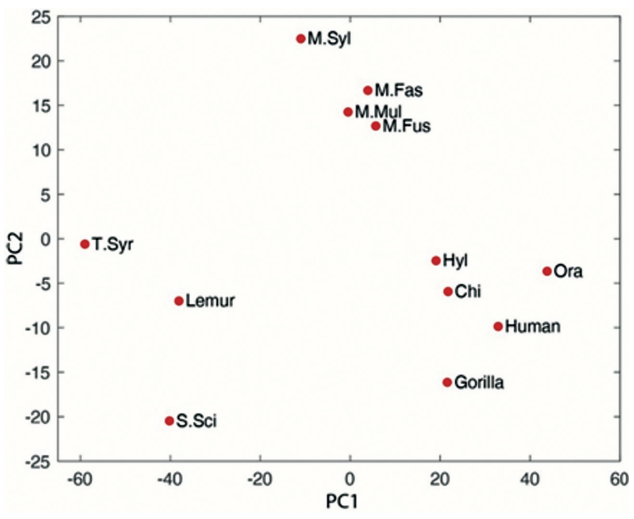


**Figure 4.** The projection of the 7-dimensional vectors of 12 species into 2D property space, which consists of two main principal components, PC1 and PC2.

**Table 2.** Whole mitochondrial genome detailed information of 18 eutherian mammals from NCBI database

| | Species | Accession Code | Length (bp) |
|---|---|---|---|
| **1** | *Human* | V00662 | 16569 |
| **2** | *Pygmy chimpanzee* | D38116 | 16563 |
| **3** | *Common chimpanzee* | D38113 | 16554 |
| **4** | *Gorilla* | D38114 | 16364 |
| **5** | *Orangutan* | D38115 | 16389 |
| **6** | *Gibbon* | X99256 | 16472 |
| **7** | *Baboon* | Y18001 | 16521 |
| **8** | *Horse* | X79547 | 16660 |
| **9** | *White rhinoceros* | Y07726 | 16832 |
| **10** | *Harbor seal* | X63726 | 16826 |
| **11** | *Gray seal* | X72004 | 16797 |
| **12** | *Cat* | U20753 | 17009 |
| **13** | *Fin whale* | X61145 | 16397 |
| **14** | *Blue whale* | X72204 | 16402 |
| **15** | *Cow* | V00654 | 16338 |
| **16** | *Rat* | X14848 | 16300 |
| **17** | *Mouse* | V00711 | 16295 |
| **18** | *Platypus* | X83427 | 17019 |

space, which consists of two main components, PC1 and PC2. When the groupings in the figure are examined, it can be seen that the results are generally compatible with the results above. The given principal components contain 87% of the total inertia of the 7-dimensional vector of this data set.

### Whole mitochondrial genomes of 18 eutherian mammals

Whole mitochondrial genomes contain wealthy genetic information and have been frequently investigated in recent years. Here, the whole mitochondrial genomes of 18 eutherian mammals were assessed [40, 41]. All of the sequences were obtained from the NCBI genetic database and are given in Table 2. The lengths of genomes ranged from 16,295 to 17,019 base pairs.

We applied the proposed method to the sequences given in Table 2 and obtained the feature vectors. We then calculated the similarities between these vectors using Euclidean distances. The phylogenetic tree obtained by this calculation is given in Figure 5. We analyzed the same DNA sequences in MEGA7 software by the alignment-based method ClustalW. We used the UPGMA method to construct the phylogenetic tree presented in Figure 6.

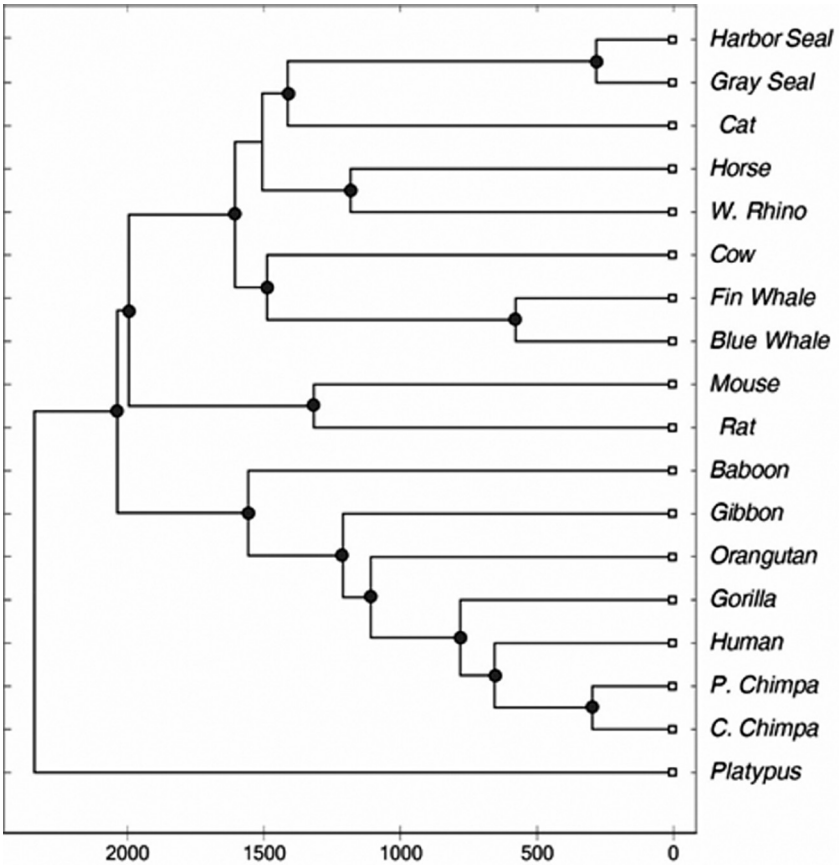


**Figure 5.** Phylogenetic tree generated by the proposed method.

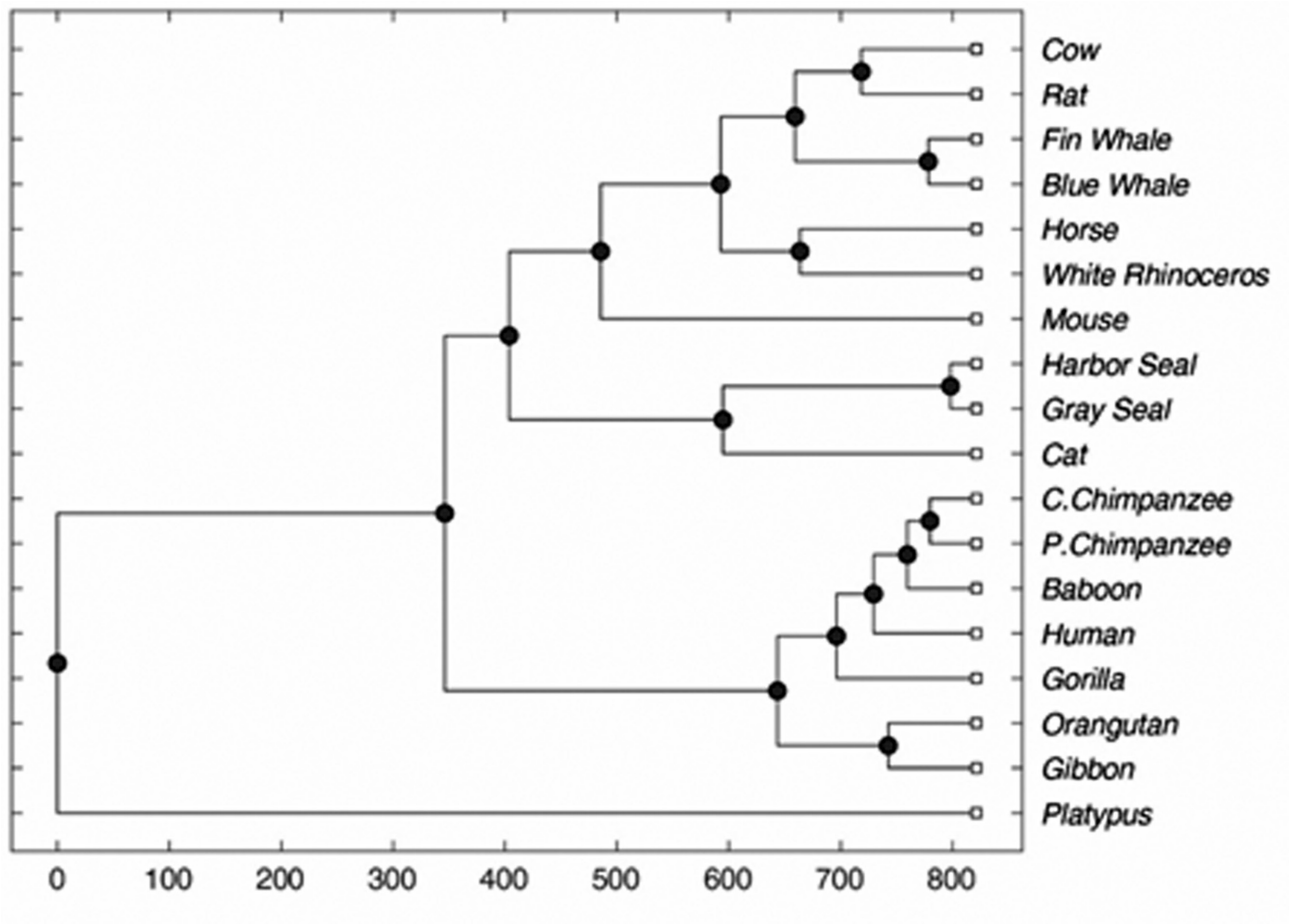


**Figure 6.** Phylogenetic tree generated by MEGA7 based on ClustalW alignment and the UPGMA method.

The trees generated by the proposed method and the reference method (MEGA7) showed overall qualitative agreement in the similarity matrix. To clarify and visualize this result, in addition, in the distance measurement plots in Figure 7, we revealed the similarity between the human and other species from the sequences given in Table 2. The compatibility of these two curves shows the general agreement between the proposed method and the results obtained with the reference method. In Figure 8, it shows the projection of eighteen feature vectors to the 2D property space, which consists of two main components, PC1 and PC2. When the groupings in the figure are examined, it can be seen that the results are generally compatible with the results above. The given principal components contain 90.82% of the total inertia of the 7-dimensional vector of this data set.

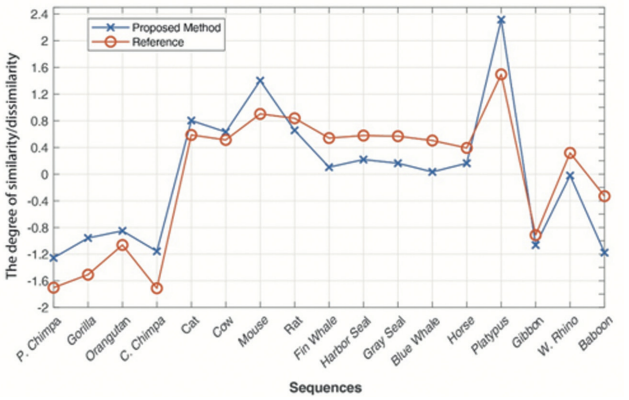


**Figure 7.** The degree of similarity between human and the other 17 species.

**Method Performance**

When selecting a DNA sequence similarity analysis method, the performance of an algorithm in terms of a reduced computational cost should be considered. DNA sequences have various lengths, ranging from approximately 100 to 20 K base pairs. The number of sequences to be compared and the sequence lengths are important factors that affect computational cost. Reducing the computational cost is a primary goal of algorithms and methods that perform alignment-free similarity analysis. One of the advantages of our proposed method is that it has

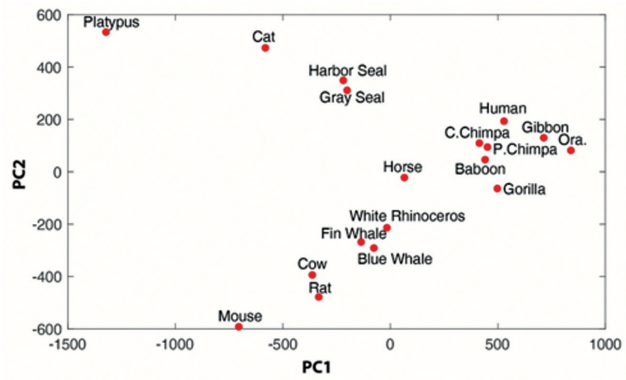


**Figure 8.** The projection of the 7-dimensional vectors of 18 species into 2D property space, which consists of two main principal components, PC1 and PC2.

low time complexity. In our method, a DNA sequence is scanned only once to generate a small size feature representation vector consisting of 7 dimensions. A vector for a single DNA sequence of length $n$ can be generated in $O(n)$ time units. If the number of the compared sequences is $m$, then all vectors for which similarity measurement is to be performed can be generated in $O(mn)$ time units. Table 3 shows the computation times of the two frequently used alignment-based methods with Mega7 tool and the computation times of the proposed method. When the table is examined, the performance of the proposed method in terms of calculation time is clearly seen.

## CONCLUSION

Here, we presented a novel feature vector model for the alignment-free numerical characterization of DNA sequences by generating a 7-dimensional vector. Our proposed method is based on mean distance of the transitions, mean distance of the nucleotide duplications, and the base frequencies. A feature vector can be obtained by using these different features of DNA sequences. In turn, the vector is characteristic of the sequence. Euclidean distances are used to calculate the similarity between feature vectors corresponding to the sequences in the dataset. The complexity cost, which is the main drawback of alignment-based methods, is mitigated by reducing the complexity of our proposed method. When considering accuracy, the dendrogram trees generated by the proposed method show the same topologies and are consistent with trees found in the literature, yet they are more successful than previous findings. There was overall qualitative agreement in the distance matrices between the trees generated by the proposed method and those of the reference method (MEGA7). To visualize and clarify this, we denoted the degree of similarity between human and other species given in the datasets by distance measurement plots. The compatibility of these two curves shows the general agreement between the proposed method and the results obtained with the reference method. These rough projections confirm that the mathematical identifier effectively characterizes the DNA sequence.

**Table 3.** Performance comparison between proposed method results and alignment-based methods.

| | Computation times (sec) | | |
|---|---|---|---|
| **Datasets** | **Proposed Method** | **ClustalW** | **Muscle** |
| Table 1 | 0.68 | 6.60 | 2.02 |
| Table 2 | 23.07 | 4528.05 | 2877.58 |

## AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

## DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

## CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## ETHICS

There are no ethical issues with the publication of this manuscript.

## REFERENCES

[1] Robbins RJ, Benton D, Snoddy J. Informatics and the Human-Genome-Project. IEEE Eng Med Biol 1995;14:694–701. [CrossRef]

[2] Wang S, Tian F, Qiu Y, Liu X. Bilateral similarity function: A novel and universal method for similarity analysis of biological sequences. J Theor Biol 2010;265:194–201. [CrossRef]

[3] Jin X, Jiang Q, Chen Y, Lee S-J, Nie R, Yao S, et al. Similarity/dissimilarity calculation methods of DNA sequences: A survey. J Mol Graph Model 2017;76:342–355. [CrossRef]

[4] Eddy SR. What is dynamic programming? Nat Biotechnolv2004;22:909–910. [CrossRef]

[5] Needleman SB, Wunsch CD. A general method applicable to search for similarities in amino acid

sequence of 2 proteins. J Mol Biol 1970;48:443–453. [CrossRef]
[6] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197. [CrossRef]
[7] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–410. [CrossRef]
[8] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85:2444–2448. [CrossRef]
[9] Hamori E, Ruskin J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. J Biol Chem 1983;258:1318–1327. [CrossRef]
[10] Nandy A. A new graphical representation and analysis of DNA-sequence structure: 1. methodology and application to globin genes. Curr Sci 1994;66:309–314.
[11] Randic M, Vracko M, Lers N, Plavsic D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Physics Lett 2003;368:14. [CrossRef]
[12] Dai Q, Liu X, Wang T. A novel 2D graphical representation of DNA sequences and its application. J Mol Graph Model 2006;25:340–344. [CrossRef]
[13] Guo Y, Wang TM. A new method to analyze the similarity of the DNA sequences. J Mol Struct-Theochem 2008;853:62–67. [CrossRef]
[14] Yu JF, Wang JH, Sun X. Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation. Match-Commun Math Comput Chem 2010;63:493–512.
[15] Liao B, Xiang QL, Cai LJ, Cao Z. A new graphical coding of DNA sequence and its similarity calculation. Phys A Stat Mech Appl 2013;392:4663–4667. [CrossRef]
[16] Hayasaka K, Gojobori T, Horai S. Molecular phylogeny and evolution of primate mitochondrial-DNA," molecular biology and evolution. 1988;5:626–644.
[17] Qi XQ, Wen J, Qi ZH. New 3D graphical representation of DNA sequence based on dual nucleotides. J Theor Biol 2007;249:681–690. [CrossRef]
[18] Yu JF, Sun X, Wang JH. TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. J Theor Biol 2009;261:459–468. [CrossRef]
[19] Jafarzadeh N, Iranmanesh A. C-curve: a novel 3D graphical representation of DNA sequence based on codons. Math Biosci 2013;241:217–224. [CrossRef]
[20] Wąż P, Bielińska-Wąż D. Non-standard similarity/dissimilarity analysis of DNA sequences. Genomics 2014;104:464–471. [CrossRef]
[21] Liao B, Tan MS, Ding KQ. A 4D representation of DNA sequences and its application. Chem Phys Lett 2005;402:380–383. [CrossRef]
[22] Liao B, Li RF, Zhu W, Xiang XY. On the similarity of DNA primary sequences based on 5-D representation. J Math Chem 2007;42:47–57. [CrossRef]
[23] Stan C, Cristescu CP, Scarlat EI. Similarity analysis for DNA sequences based on chaos game representation. Case study: the albumin. J Theor Biol 2010;267:513–518. [CrossRef]
[24] Hoang T, Yin C, Yau SS. Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. Genomics 2016;108:134–142. [CrossRef]
[25] Qi X, Wu Q, Zhang Y, Fuller E, Zhang CQ. A novel model for DNA sequence similarity analysis based on graph theory. Evol Bioinform Online 2011;7:149–158. [CrossRef]
[26] Rinku M, Neeru A. A graph theoretic model for prediction of reticulation events and phylogenetic networks for DNA sequences. Egyptian J Basic Appl Sci 2016;3:263–271. [CrossRef]
[27] Zhou J, Zhong P, Zhang T. A novel method for alignment-free DNA sequence similarity analysis based on the characterization of complex networks. Evol Bioinform Online 2016;12:229–235. [CrossRef]
[28] Liu L, Ho YK, Yau S. Clustering DNA sequences by feature vectors. Mol Phylogenet Evol 2006;41:64–69. [CrossRef]
[29] Wu RH, Hu QG, Li RF, Yue GX. A novel composition coding method of DNA Sequence and its application. Match-Commun Math Comput Chem 2012;67:269–276.
[30] Qi X, Fuller E, Wu Q, Zhang CQ. Numerical characterization of DNA sequence based on dinucleotides. ScientificWorldJournal 2012;2012:104–269. [CrossRef]
[31] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol 2017;18:186. [CrossRef]
[32] Zhang YS. A simple method to construct the similarity matrices of DNA sequences. Match-Commun Math Comput Chem 2008;60:313–324.
[33] Shi L, Huang HL. DNA sequences analysis based on classifications of nucleotide bases. Affect Comput Intel Inter 2012;137:379–384. [CrossRef]
[34] Bao J, Yuan R, Bao Z. An improved alignment-free model for DNA sequence similarity metric. BMC Bioinformatics 2014;15:321. [CrossRef]
[35] Kumar R, Mishra BK, Lahiri T, Kumar G, Kumar N, Gupta R, et al. PCV: An Alignment Free Method for Finding Homologous Nucleotide Sequences and its Application in Phylogenetic Study. Interdiscip Sci 2017;9:173–183. [CrossRef]
[36] Blackburn GM, Gait MJ, Loakes D, Williams DM. Nucleic Acids in Chemistry and Biology: Edition 3. London: Royal Society of Chemistry; 2006. [CrossRef]

[37] Chen WY, Liao B, Li WW. Use of image texture analysis to find DNA sequence similarities. J Theor Biol 2018;455:1–6. [CrossRef]

[38] Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 2016;33:1870–1874. [CrossRef]

[39] Zhang YS, Chen W. New invariant of DNA sequences. Match-Commun Math Comput Chem 2007;58:197–208.

[40] Delibaş E, Arslan A. DNA sequence similarity analysis using image texture analysis based on first-order statistics. J Mol Grap Model 2020;99:107–603. [CrossRef]

[41] Jin X, Nie RC, Zhou DM, Yao SW, Chen YY, Yu JF, et al. A novel DNA sequence similarity calculation based on simplified pulse-coupled neural network and Huffman coding. Phys A Stat Mech Appl 2016;461:325–338. [CrossRef]