MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **97** (2027) 941–960
ISSN: 0340–6253
doi: 10.46793/match.97-3.11326

A New Alignment-Free Method for the Phylogenetic Analysis of DNA Sequences

En Wu, Jie Feng*

School of Science, Minzu University of China, Beijing 100081, China
18702048279@163.com, fengjie0536@163.com

(Received April 2, 2026)

Abstract

This paper defines the Accumulated Nucleotide Frequency Distribution (ANFD) of DNA sequences. We compute the mean, variance, and covariance of the ANFD for each base, together with the mean of the Accumulated Nucleotide Frequency (ANF). Based on these quantities, a 14-dimensional feature vector is constructed to numerically represent DNA sequences. The Euclidean distance is then employed to quantify the similarity between pairs of DNA sequences. The proposed method is validated experimentally on three DNA sequence datasets. The results demonstrate that the method achieves accurate clustering of DNA sequences, thus verifying its effectiveness.

1 Introduction

With the rapid development of high-throughput sequencing technologies, genomic data from an increasing number of species have become available. Accurately inferring evolutionary relationships from such large-scale data remains a major challenge. Traditional phylogenetic methods mainly rely

*Corresponding author: fengjie0536@163.com

This work is licensed under a Creative Commons "Attribution 4.0 International" license.

on multiple sequence alignment (MSA), which uses homologous sites identified through alignment to reconstruct evolutionary relationships. However, MSA faces limitations when applied to large datasets, distantly related species, and genomes or non-coding regions with frequent rearrangements. These limitations include high computational cost, low efficiency, accumulation of alignment errors, and strong reliance on homology assumptions, which may result in biased or inaccurate phylogenetic inference.

Alignment-free methods have gained increasing attention in recent years as useful complements or alternatives to traditional phylogenetic methods. These methods do not use sequence alignment. Instead, they describe sequences using overall features such as k-mer frequency, information entropy, compression complexity, and word frequency distance to measure similarity between species. They are efficient in computation and are suitable for short sequences or sequences without clear homology. Therefore, they work well for whole-genome data, large non-coding regions, and comparisons between distantly related species.

Over the past few decades, researchers have used various alignment-free methods to classify biological sequences. In these methods, the key step is to extract feature vectors for sequence comparison. Common feature extraction approaches include the k-mer method, image representation, and information theory–based methods.

(1) k-mer method

The k-mer method, proposed by Blaisdell [3] in 1985, is one of the most widely used alignment-free methods. It represents sequences based on the frequency of k-length substrings. For DNA sequences, there are 4^k possible k-mer combinations for a given k. For example, when $k = 2$, all substrings of length 2 are considered. The full set of 2-mers is : {AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT}, with a total of $4^2 = 16$ combinations. By counting the frequency of each 2-mer in a DNA sequence, we obtain a 16-dimensional feature vector.

However, simple frequency features ignore the position of nucleotides or amino acids in a sequence. To address this, Chou et al. [7] proposed pseudo-amino acid composition. This method keeps amino acid frequency

information and also adds correlation between nearby amino acids. These correlations are usually based on physicochemical properties within a certain range. The pseudo-amino acid composition has since been widely used in protein sequence analysis [30, 34, 35, 37, 38]. In addition, Wang et al. [39] studied how the choice of k affects sequence similarity results. In practice, the k-mer method is often combined with other statistical features to build feature vectors, which helps reduce the loss of sequence information [1, 4, 9, 36, 40–42].

(2) Image representation

Graphical representation of biological sequences is an alignment-free method widely used for similarity analysis of nucleotide or amino acid sequences. In 1983, Hamori and Ruskin [14] first proposed the H-curve, which represents DNA sequences in 3D space. Later, Gates [11], Nandy [31], and Leong and Morgenthaler [20] proposed simpler 2D curve representations. Academician Zhang Chunting [43, 44] introduced the Z-curve, a classic 3D method for representing DNA sequences. The H-curve can be seen as a special case of the Z-curve. From the perspective of the Z-curve, the H-curve is simply a symmetrical, fragmented Z-curve. Other curves, such as four-color curves and zigzag curves, can also be viewed as variations of the Z-curve, showing that it provides a general framework for this type of method.

In 2006, Liao and Ding et al. [23] proposed a new 3D representation that avoids information loss caused by curve overlap in earlier methods. In 2008, Cao et al. [5] further simplified DNA sequences into four 3D representations based on the chemical properties of adjacent dinucleotides, and combined them with covariance matrix features to measure sequence similarity. In addition, researchers have proposed more image representation methods [2, 6, 8, 17, 24–26, 32, 45, 46] .

(3) Information theory method

As research has progressed, information theory methods have been introduced for biological sequence analysis. These methods measure sequence differences by estimating how much information is shared between sequences. Among them, complexity and entropy are the most commonly used tools.

In 2001, complexity was first applied to similarity analysis. Li et al. [21] used compression algorithms to estimate sequence complexity, and then computed conditional complexity through further compression. Based on these values and a distance formula, they quantified similarity between biological sequences. However, the results were only approximate. To improve this, Lempel and Ziv [19] proposed Lempel–Ziv complexity, which provides a more accurate and effective way to analyze sequence similarity.

Entropy is also widely used in biological sequence analysis to measure uncertainty in a system. Gatlin [12] proposed the conditional entropy method; in 1948, Shannon [33] introduced Shannon entropy to address the quantitative measurement of information. Hariri et al. [15] proposed the k-Shannon entropy method based on Shannon entropy; Fang [10] proposed a method for measuring information discreteness, among others. In recent years, with the continuous development and improvement of information theory, information theory methods have been increasingly applied by researchers in the field of bioinformatics [18, 22, 27, 29, 47–49].

In addition to the three common methods mentioned above, researchers have also discovered many novel feature extraction methods. For example, He et al. [16] proposed the Correlation Coefficient Eigenvector (CCFV) to extract the distribution characteristics of nucleotide positions in DNA sequences for evolutionary analysis of common human viruses. Chaos Game Representation (CGR) is a milestone in graphic bioinformatics and has become an incredibly powerful tool for sequence comparison and feature encoding in machine learning [28]. Muhammed et al. [13] used genome mapping technology to convert genome sequences into grayscale images, and obtained several statistical features from the images for training and testing multiple classifiers. Kshatrapal et al. [50] divided four nucleotide bases into two groups based on their chemical properties, resulting in a total of three classifications. Three symbol sequences were constructed for a DNA sequence, and the frequency of intra group mutations was calculated for each symbol sequence. Finally, a 12 dimensional vector was constructed to represent a sequence. For more methods, please refer to the review literature [51–54].

Despite these advances, further improvements in the quantitative char-

acterization of DNA sequences are still highly desirable. In the present work, we innovatively propose the definition of the Accumulated Nucleotide Frequency Distribution (ANFD), extract a set of numerical features pertaining to base-specific ANFD metrics, and further establish a 14-dimensional feature vector for the numerical representation of DNA sequences. To eliminate the interference of data value ranges on subsequent analytical outcomes, standardization processing is implemented on the constructed feature vector. On the basis of the standardized vector representations of individual sequences, the Euclidean distance is adopted to quantitatively characterize the inter-sequence similarity. The validity and feasibility of the proposed method are comprehensively validated through experiments on multiple independent datasets.

2 Constructing the feature vector of DNA sequence

2.1 Accumulated nucleotide frequency (ANF)

Let a DNA sequence of length n be denoted by $S = s_1 s_2 \cdots s_n$, where s_i represents the nucleotide at position i. The Accumulated Nucleotide Frequency (ANF) [55] is defined as follows: In sequence S, the accumulated nucleotide frequency $f_b(k)$ of the k-th base b is the ratio between the cumulative number of occurrences of base b before (including) the current position and its current position in the sequence. Formally, it can be expressed as follows.

$$f_b(k) = \frac{k}{L_b(k)}, \quad k = 1, 2, \ldots, n_b \tag{1}$$

where $L_b(k)$ represents the position where the k-th base b appears in sequence S; n_b is the total number of occurrences of base b in the sequence.

This frequency reflects the local enrichment trend of a certain type of base in the entire sequence, that is, its historical accumulated density. Taking the sequence $S = ACTGGCAAT$ as an example, the accumulated nucleotide frequencies of each base in the sequence are as follow:

$\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{2}{5}, \frac{2}{6}, \frac{2}{7}, \frac{3}{8}, \frac{2}{9}$.

2.2 Accumulated nucleotide frequency distribution (ANFD)

Let S be a DNA sequence of length n. For a specific base $b \in \{A, T, C, G\}$, the calculation rule for the accumulated nucleotide frequency distribution $U_b(i)$ at the i-th position of base b is:

When $L_b(1) \neq 1$, that is, base b does not appear at the first position of sequence S, the accumulated nucleotide frequency distribution of base b is:

$$U_b(i) = \begin{cases} 0, & 1 \leq i < L_b(1), \\ \dfrac{f_b(k)}{L_b(k+1) - L_b(k)}, & L_b(k) \leq i < L_b(k+1);\ 1 \leq k < n_b, \\ \dfrac{f_b(n_b)}{n - L_b(n_b) + 1}, & L_b(n_b) \leq i \leq n. \end{cases} \quad (2)$$

When $L_b(1) = 1$, that is, base b appears at the first position of sequence S, the accumulated nucleotide frequency distribution of base b is:

$$U_b(i) = \begin{cases} \dfrac{f_b(k)}{L_b(k+1) - L_b(k)}, & L_b(k) \leq i < L_b(k+1);\ 1 \leq k < n_b, \\ \dfrac{f_b(n_b)}{n - L_b(n_b) + 1}, & L_b(n_b) \leq i \leq n. \end{cases} \quad (3)$$

Taking the sequence $S = ACTGGCAAT$ as an example (n=9): The positions where A appears are: $\{1, 7, 8\}$, $n_A = 3$. When $1 \leq i < 7$, $U_A(i) = \frac{\frac{1}{1}}{7-1} = \frac{1}{6}$; when $7 \leq i < 8$, $U_A(i) = \frac{\frac{2}{7}}{8-7} = \frac{2}{7}$; when $8 \leq i \leq 9$, $U_A(i) = \frac{\frac{3}{8}}{9-8+1} = \frac{3}{16}$. The same applies to other bases. The accumulated nucleotide frequency distribution of the sequence $S = ACTGGCAAT$ is shown in Table 1:

Table 1. The accumulated nucleotide frequency distribution of the sequence $S = ACTGGCAAT$

sequence	A	C	T	G	G	C	A	A	T
$U_A(i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{7}$	$\frac{3}{16}$	$\frac{3}{16}$
$U_C(i)$	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
$U_T(i)$	0	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{2}{9}$
$U_G(i)$	0	0	0	$\frac{1}{4}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{25}$

2.3 Statistical features

Based on the ANFD of the sequence, the mean, variance, and covariance of the ANFD for each nucleotides A, T, C, and G, as well as the mean of the ANF, can be calculated.

The mean of the ANFD of each nucleotides is denoted as $\overline{U}_b$, and its calculation formula is:

$$\overline{U}_b = \frac{\sum_{i=1}^{n} U_b(i)}{n}, \quad b \in \{A, T, C, G\} \tag{4}$$

The covariance formula for the ANFD between nucleotides is:

$$\mathrm{Cov}(b_1, b_2) = \frac{\sum_{i=1}^{n} \left(U_{b_1}(i) - \overline{U}_{b_1}\right)\left(U_{b_2}(i) - \overline{U}_{b_2}\right)}{n_{b_1} \cdot n_{b_2}} \tag{5}$$

Where $b_1, b_2 \in \{A, T, C, G\}$, $b_1 \neq b_2$, n_{b_1} and n_{b_2} represent the number of occurrences of nucleotide b_1 and n_{b_2} in the sequence, respectively.

According to the covariance formula, when $b_1 = b_2$, covariance is equivalent to variance:

$$D_b = \frac{\sum_{i=1}^{n} \left(U_b(i) - \overline{U}_b\right)^2}{n_b^2}, \quad b \in \{A, T, C, G\} \tag{6}$$

The mean value of the accumulated nucleotide frequency of each nucleotides is denoted as $\overline{K}_b$, which is defined as:

$$\overline{K}_b = \frac{\sum_{i=1}^{n} U_b(i)}{n_b}, \quad b \in \{A, T, C, G\} \tag{7}$$

2.4 14-dimensional feature vector of DNA sequence

The mean value of the ANF for each nucleotides, denoted as $\overline{K}_b$, the variance of the ANFD, denoted as D_b, and the $\mathrm{Cov}(b_1, b_2)$, are constructed into a 14-dimensional feature vector to numerically represent DNA sequences. The 14-dimensional feature vector corresponding to the i-th sequence is denoted as: $L_i = \big(\overline{K}_{Ai},\ \overline{K}_{Ci},\ \overline{K}_{Gi},\ \overline{K}_{Ti},\ D_{Ai},\ D_{Ci},\ D_{Gi},\ D_{Ti},$ $\mathrm{Cov}_i(A,C),\ \mathrm{Cov}_i(A,T),\ \mathrm{Cov}_i(C,G),\ \mathrm{Cov}_i(C,T),\ \mathrm{Cov}_i(G,T)\big)$. To eliminate the influence of inconsistent dimensions, the feature vector needs to be standardized first.

3 Similarity and evolutionary analysis of DNA sequences

To validate the effectiveness of the method proposed in this study, we apply it to three representative datasets: β-globin gene sequences from 11 species, 59 hepatitis C virus gene sequences, and 20 complete mammalian mitochondrial genome sequences. The Euclidean distance is adopted to compute pairwise distances between the corresponding feature vectors, and the UPGMA algorithm implemented in MEGA11 is used to reconstruct phylogenetic trees. All sequences in the three datasets were retrieved from the NCBI database.

3.1 Sequences of β-globin gene from 11 species

In this section, β-globin sequences from 11 species are selected for experimental validation, with detailed sequence information listed in Table 2. Using the proposed method, an 11×14 feature matrix is constructed for the 11 β-globin sequences. Pairwise Euclidean distances are then computed to generate a distance matrix, as displayed in Table 3. A smaller distance between two feature vectors indicates higher evolutionary similarity between the corresponding species.

As observed in Table 3, the distance between Chimpanzee and Gorilla is the smallest (0.8143), reflecting their closest evolutionary relationship.

The distances between Human and Lemur, Gorilla, and Chimpanzee are 1.4994, 1.4638, and 1.4468, respectively, all of which are relatively small, indicating a high degree of similarity among these four species. In contrast, Gallus exhibits relatively large distances to all other species, implying a distant evolutionary divergence, which is consistent with established biological knowledge and known evolutionary relationships.

Table 2. β-globin sequence information of 11 species

Species	Accession	Length (bp)
Human	U01317	1424
Goat	M15387	1471
Opossum	J03643	2022
Gallus	V00409	1346
Lemur	M15734	1442
Mouse	V00722	1188
Rat	X06701	1196
Gorilla	X61109	1344
Bovine	X00376	1464
Chimpanzee	X02345	1344
Rabbit	V00878.1	1399

Table 3. Similarity Distance Matrix of β - globin Gene Sequence in 11 Species

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rat	Gorilla	Bovine	Chimpanzee	
Goat	3.2514										
Opossum	6.5036	4.6573									
Gallus	9.2865	7.4181	9.6527								
Lemur	1.4994	3.3672	6.9636	9.0488							
Mouse	4.7511	5.0788	8.0607	7.7384	4.1807						
Rat	4.4915	4.5241	7.1440	8.5519	4.1250	2.7544					
Gorilla	1.4638	3.9413	7.2608	9.5043	2.1408	4.3710	3.9949				
Bovine	3.2128	1.3482	5.4875	7.2278	3.1280	4.3000	3.9834	3.5753			
Chimpanzee	1.4468	3.9448	7.3123	9.3183	2.0250	4.1823	4.1528	0.8143	3.6547		
Rabbit	4.0265	5.3604	8.6393	9.8297	3.7205	3.9204	2.6036	3.2076	4.8914	3.4498	

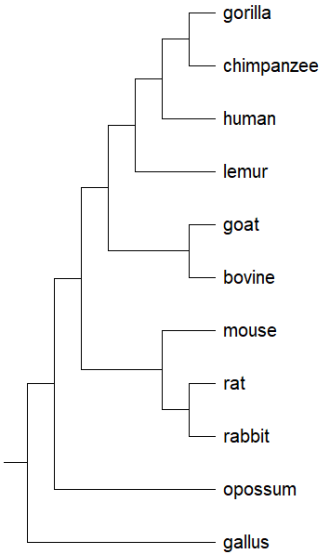

Figure 1. The phylogenetic tree of β-globin sequences of 11 species constructed by our method

Figure 1 displays the phylogenetic tree constructed for the β-globin sequences of the 11 species using the proposed method. Visual inspection shows that the sequences are clearly clustered into three major groups: primates (gorilla, chimpanzee, human, lemur), artiodactyla (goat, bovine), and rodents (rat, mouse). Primates and artiodactyla exhibit a relatively close evolutionary relationship, whereas Gallus, as the sole non-mammalian species, is distinctly separated from the others and occupies an outermost branch in the tree. These results are fully consistent with known biological evolutionary relationships.

3.2 59 types of hepatitis C virus (HCV) gene sequences

Detailed sequence information for the 59 hepatitis C virus isolates is summarized in Table 4. The corresponding phylogenetic tree constructed using the proposed method is shown in Figure 2. As can be observed, genotypes 1, 2, 3, and 5 are correctly classified into their respective clades.

Table 4. 59 types of hepatitis C virus (HCV) sequences information

Accession	Types	Length (bp)
JQ717260	type3	9425
KF035123	type3	9429
GQ275355	type3	9442
KY620860	type3	9449
KY620677	type3	9422
KY620662	type3	9443
KY620679	type3	9432
KY620833	type3	9446
KY620681	type3	9455
KY620473	type3	9416
KY620782	type3	9442
KY620710	type3	9446
KY620828	type3	9421
KY620829	type3	9454
KY620831	type3	9422
JQ717257	type3	9425
KF035124	type3	9426
KY620664	type3	9421
JN714194	type3	9429
JQ717255	type3	9425
EF407478	type1	9376
EU862837	type1	9336
EU155220	type1	9316
EU256000	type1	9329
EU255961	type1	9332
EU155369	type1	9330
EU155328	type1	9329
EU155337	type1	9329
EU155370	type1	9329
EU155261	type1	9330

Continued on next page

Accession	Types	Length (bp)
EU482886	type1	9333
EU256092	type1	9329
EU155308	type1	9329
KU180728	type1	9374
EU155302	type1	9329
HB455523	type2	9666
HB455528	type2	9666
HB455524	type2	9666
HB455527	type2	9666
HB455526	type2	9666
HB455525	type2	9666
HB455597	type2	9666
HC510319	type2	9665
HQ852455	type2	9666
HQ852456	type2	9666
HM049503	type2	9666
HQ852454	type2	9666
HB455538	type2	9666
HB455539	type2	9666
HB455598	type2	9666
HB455540	type2	9666
HB455599	type2	9666
HB455600	type2	9666
JX227973	type4	9092
MH119606	type4	9027
JE984232	type5	9620
JA212565	type5	9621
JA212566	type5	9621
JA212567	type5	9621

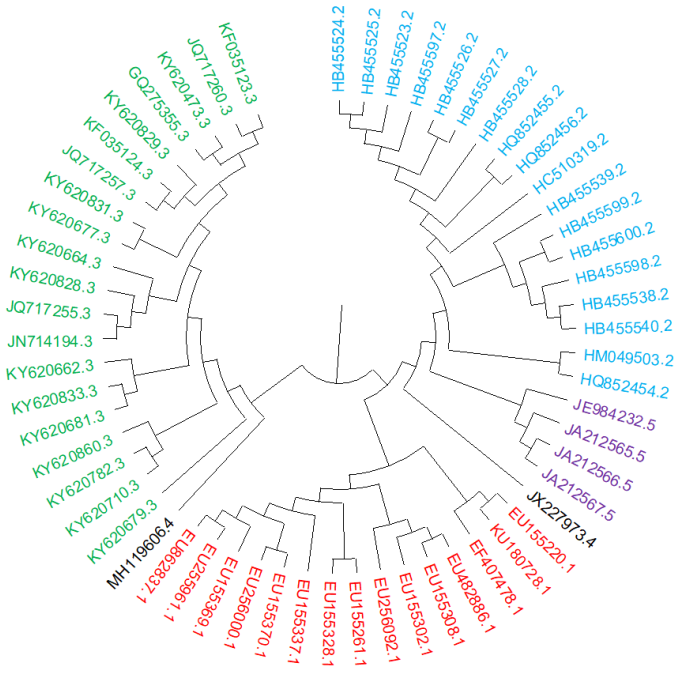

Figure 2. The phylogenetic tree of 59 hepatitis C virus gene sequences constructed by our method

3.3 Sequence of the complete mitochondrial genomes of 20 placental mammals

Sequence information for the complete mitochondrial genomes of 20 placental mammals is listed in Table 5. The phylogenetic tree reconstructed using the present method (Figure 3) indicates that the species can be broadly divided into the clade of Rodentia and a combined clade of Primates and Perissodactyla. This topology confirms a closer evolutionary affinity between Primates and Perissodactyla.

Table 5. β-globin sequence information of 11 species

Species	Accession	Length (bp)
human	V00662	16569
pigmy chimpanzee	D38113	16554
orangutan	D38115	16389
baboon	Y18001	16521
white rhinoceros	Y07726	16832
gray seal	X72004	16797
fin whale	X61145	16398
cow	V00654	16338
mouse	V00711	16295
wallaroo	Y10524	16896
common chimpanzee	D38116	16563
gorilla	D38114	16364
gibbon	X99256	16472
horse	X79547	16660
harbor seal	X63726	16826
cat	U20753	17009
blue whale	X72204	16402
rat	X14848	16300
opossum	Z29573	17084
platypus	X83427	17019

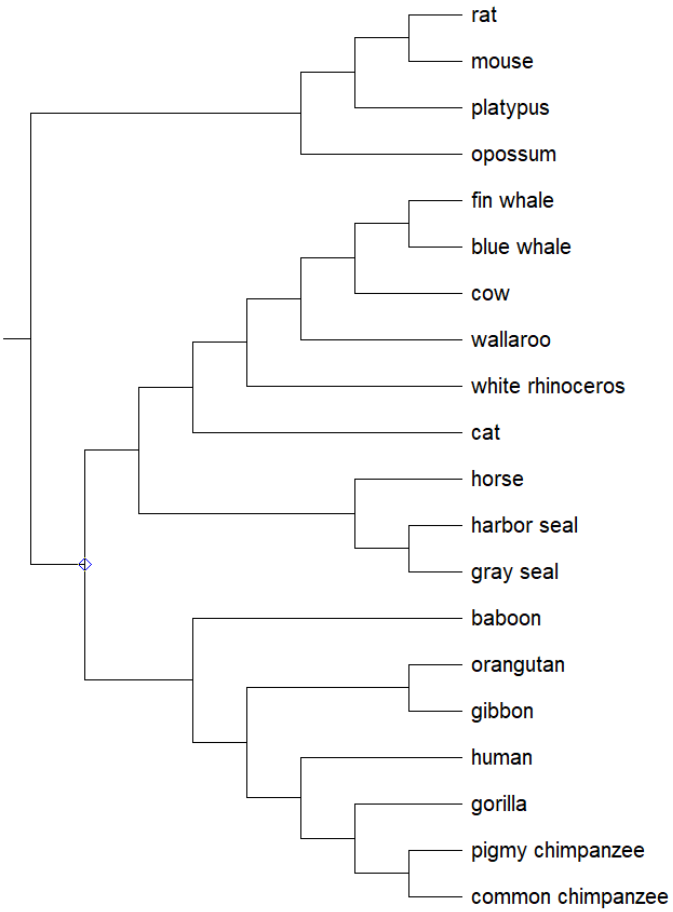

Figure 3. The phylogenetic tree of 20 mammalian mitochondrial whole genome sequences constructed by our method

4 Conclusion

This paper presents a novel alignment-free computational method that introduces the concept of the accumulated nucleotide frequency distribution of individual bases. Using this distribution, we compute the mean accumulated frequencies of the four nucleotides, the variances of their accumulated frequency distributions, and the pairwise covariances between

bases. In this way, each DNA sequence is represented as a 14-dimensional numerical vector. The proposed method yields accurate and biologically consistent clustering results across three benchmark datasets, demonstrating its reliability and effectiveness for DNA sequence similarity analysis.

References

[1] L. Ai, J. Feng, Y. H. Yao, A novel fast approach for protein classification and evolutionary analysis, *MATCH Commun. Math. Comput. Chem.* **90** (2023) 381–398.

[2] D. Bielińska-Wąż, Graphical and numerical representations of DNA sequences: Statistical aspects of similarity, *J. Math. Chem.* **49** (2011) 2345–2407.

[3] D. Bielińska-Wąż, P. Wąż, D. Panas, Applications of 2D and 3D-dynamic representations of DNA/RNA sequences for a description of genome sequences of viruses, *Comb. Chem. High Throughput Screen.* **25** (2022) 429–438.

[4] B. E. Blaisdell, Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding, *J. Mol. Evol.* **21** (1985) 278–288.

[5] V. Bonnici, A. Cracco, G. Franco, A k-mer based sequence similarity for pangenomic analyses, in: *International Conference on Machine Learning, Optimization, and Data Science*, Springer, Cham, 2021, pp. 31–44.

[6] Z. Cao, B. Liao, R. F. Li, A group of 3D graphical representation of DNA sequences based on dual nucleotides, *Int. J. Quantum Chem.* **108** (2008) 1485–1490.

[7] R. Chi, K. Q. Ding, Novel 4D numerical representation of DNA sequences, *Chem. Phys. Lett.* **407** (2005) 63–67.

[8] K. C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, *Proteins Struct. Func. Bioinf.* **43** (2010) 246–255.

[9] E. Delibaş, A. Arslan, DNA sequence similarity analysis using image texture analysis based on first-order statistics, *J. Mol. Graph. Model.* **99** (2020) #107603.

[10] E. Delibaş, A. Arslan, A. Şeker, B. Diri, A novel alignment-free DNA sequence similarity analysis approach based on top-k n-gram match-up, *J. Mol. Graph. Model.* **100** (2020) #107693.

[11] W. W. Fang, F. S. Roberts, Z. R. Ma, A measure of discrepancy of multiple sequences, *Inf. Sci.* **137** (2001) 75–102.

[12] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol.* **119** (1986) 319–328.

[13] L. Gatlin, *Information Theory and the Living System*, Columbia Univ. Press, New York, 1978.

[14] M. S. Hammad, M. S. Mabrouk, W. I. Al-atabany, V. F. Ghoneim, Genomic image representation of human coronavirus sequences for COVID-19 detection, *Alex. Eng. J.* **63** (2023) 583–597.

[15] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.* **258** (1983) 1318–1327.

[16] A. Hariri, B. Weber, J. O. Rd, On the validity of Shannon-information calculations for molecular biological sequences, *J. Theor. Biol.* **147** (1990) 235–254.

[17] L. L. He, S. Y. Sun, Q. Y. Zhang, X. N. Bao, P. K. Li, Alignment-free sequence comparison for virus genomes based on location correlation coefficient, *Infect. Genet. Evol.* **96** (2021) #105106.

[18] Y. J. Huang, T. M. Wang, New graphical representation of a DNA sequence based on the ordered dinucleotides and its application to sequence analysis, *Int. J. Quantum Chem.* **112** (2012) 1746–1757.

[19] K. Karmeshu, A. Krishnamachari, Sequence variability and long-range dependence in DNA: An information theoretic perspective, *Lect. Notes Comput. Sci.* **3316** (2004) 1354–1361.

[20] A. Lempel, J. Ziv, On the complexity of finite sequences, *IEEE Trans. Inf. Theory* **22** (1976) 75–81.

[21] P. M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Appl. Biosci.* **11** (1995) 503–507.

[22] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, H. Y. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics* **17** (2001) #149.

[23] C. Li, J. Wang, Relative entropy of DNA and its application, *Physica A* **347** (2005) 465–471.

[24] B. Liao, K. Q. Ding, A 3D graphical representation of DNA sequences and its application, *Theor. Comput. Sci.* **358** (2006) 56–64.

[25] B. Liao, R. F. Li, W. Zhu, X. Y. Xiang, On the similarity of DNA primary sequences based on 5-D representation, *J. Math. Chem.* **42** (2007) 47–57.

[26] B. Liao, T. M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases, *J. Chem. Inf. Comput. Sci.* **44** (2004) 1666–1670.

[27] H. Liu, 2D graphical representation of DNA sequence based on horizon lines from a probabilistic view, *Biosci. J.* **34** (2018) 744–750.

[28] D. M. Loewenstern, P. N. Yianilos, Significantly lower entropy estimates for natural DNA sequences, *J. Comput. Biol.* (1996) 151–161.

[29] H. F. Löchel, D. Heider, Chaos game representation and its applications in bioinformatics, *Comput. Struct. Biotech. J.* **19** (2021) 6263–6271.

[30] G. J. Ma, L. J. Liang, Y. H. Fan, W. J. Wang, J. Q. Dai, Z. F. Yuan, The entropy characters of point mutation, *Chin. Sci. Bull.* **53** (2008) 3008–3015.

[31] S. Nakariyakul, Z. P. Liu, L. N. Chen, A sequence-based computational approach to predicting PDZ domain-peptide interactions, *Biochim. Biophys. Acta* **1844** (2014) 165–170.

[32] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. methodology and application to globin genes, *Curr. Sci.* **66** (1994) 309–314.

[33] D. Q. N. Nguyen, L. Xing, P. D. T. Le, L. Lin, A graph-theoretical approach to DNA similarity analysis, *Commun. Inf. Syst.* **22** (2022) 383–400.

[34] N. Ramanathan, J. Ramamurthy, G. Natarajan, Numerical characterization of DNA sequences for alignment-free sequence comparison – A review, *Comb. Chem. High Throughput Screen.* **25** (2022) 365–380.

[35] M. Randić, M. Novič, D. Plavšić, Milestones in graphical bioinformatics, *Int. J. Quantum Chem.* **113** (2013) 2413–2446.

[36] C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27** (1948) 623–656.

[37] K. Singh, L. Singh, V. Shukla, Y. K. Sharma, A. K. Rai, An advanced approach for DNA sequencing and similarities analysis on the basis of groupings of nucleotide bases, *Int. J. Data Min. Bioinform.* **29** (2025) 133–149.

[38] S. S. Sitanshu, P. Ganapati, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction, *Comput. Biol. Chem.* **34** (2010) 246–255.

[39] W. Su, X. Q. Xie, X. W. Liu, D. Gao, Y. W. Li, iRNA-ac4C: A novel computational method for effectively detecting N4-acetylcytidine sites in human mRNA, *Int. J. Biol. Macromol.* **227** (2023) 1174–1181.

[40] P. Tripathi, P. N. Pandey, A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition, *J. Theor. Biol.* **424** (2017) #49.

[41] M. Uddin, M. K. Islam, M. R. Hassan, F. Jahan, J. H. Baek, A fast and efficient algorithm for DNA sequence similarity identification, *Complex Intell. Syst.* **9** (2023) 1265–1280.

[42] S. Vinga, J. Almeida, Alignment-free sequence comparison—a review, *Bioinformatics* **19** (2003) 513–523.

[43] S. Vinga, J. S. Almeida, Rényi continuous entropy of DNA sequences, *J. Theor. Biol.* **231** (2005) 377–388.

[44] Y. Wang, X. Y. Lei, S. Wang, Z. C. Wang, N. F. Song, F. Zeng, T. Chen, Effect of k-tuple length on sample-comparison with high-throughput sequencing data, *Biochem. Biophys. Res. Commun.* **469** (2016) 1021–1027.

[45] S. J. Wang, J. Yang, K. C. Chou, Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition, *J. Theor. Biol.* **242** (2006) 941–946.

[46] R. X. Wu, W. J. Liu, Y. Y. Mao, Z. J. Zheng, 2D graphical representation of DNA sequences based on variant map, *IEEE Access* **8** (2020) 173755–173765.

[47] M. Xiao, Z. Z. Zhu, J. P. Liu, C. Y. Zhang, A New Method Based on Entropy Theory for Genomic Sequence Analysis, *Acta Biotheor.* **50** (2002) 155–165.

[48] H. Xu, Geometric feature of DNA sequences, *Recent Pat. Eng.* **18** (2024) 86–99.

[49] W. L. Yang, X. J. Pi, L. Q. Zhang, Similarity analysis of DNA sequences based on the relative entropy, *Adv. Nat. Comput.* **3610** (2005) 1035–1038.

[50] X. W. Yang, T. M. Wang, Linear regression model of short k-word: A similarity distance suitable for biological sequences with various lengths, *J. Theor. Biol.* **337** (2013) 61–70.

[51] X. W. Yang, T. M. Wang, A novel statistical measure for sequence comparison on the basis of k-word counts, *J. Theor. Biol.* **318** (2013) 91–100.

[52] L. P. Yang, X. D. Zhang, H. G. Zhu, Alignment free comparison: K word voting model and its applications, *J. Theor. Biol.* **335** (2013) 276–282.

[53] B. Yu, S. Li, C. Chen, J. M. Xu, W. Y. Qiu, X. Wu, R. X. Chen, Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition, *Chemom. Intell. Lab. Syst.* **167** (2017) 102–112.

[54] C. T. Zhang, R. Zhang, Analysis of distribution of bases in the coding sequences by a diagrammatic technique, *Nucleic Acids Res.* **19** (1991) 6313–6317.

[55] R. Zhang, C. T. Zhang, Z curves - an intutive tool for visualizing and analyzing the DNA sequences, *J. Biomol. Struct. Dyn.* **11** (1994) 767–782.