Journal of Molecular Graphics and Modelling 99 (2020) 107603

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

DNA sequence similarity analysis using image texture analysis based on first-order statistics

Emre Delibaş [a, *], Ahmet Arslan [b]

[a] Department of Computer Engineering, Faculty of Engineering, Cumhuriyet University, 58140, Sivas, Turkey
[b] Department of Computer Engineering, Faculty of Engineering, Selçuk University, 42250, Konya, Turkey

ARTICLE INFO

Article history:
Received 17 July 2019
Received in revised form
13 March 2020
Accepted 23 March 2020
Available online 3 May 2020

Keywords:
DNA sequence similarity
Texture analysis
Alignment-free comparison

ABSTRACT

Similarity is one of the key processes of DNA sequence analysis in computational biology and bioinformatics. In nearly all research that explores evolutionary relationships, gene function analysis, protein structure prediction and sequence retrieving, it is necessary to perform similarity calculations. One major task in alignment-free DNA sequence similarity calculations is to develop novel mathematical descriptors for DNA sequences. In this paper, we present a novel approach to DNA sequence similarity analysis studies using similarity calculations of texture images. Texture analysis methods, which are a subset of digital image processing methods, are used here with the assumption that these calculations can be adapted to alignment-free DNA sequence similarity analysis methods. Gray-level textures were created by the values assigned to the nucleotides in the DNA sequences. Similarity calculations were made between these textures using histogram-based texture analyses based on first-order statistics. We obtained texture features for 3 different DNA data sets of different lengths, and calculated the similarity matrices. The phylogenetic relationships revealed by our method shows our trees to be similar to the results of the MEGA software, which is based on sequence alignment. Our findings show that texture analysis metrics can be used to characterize DNA sequences.

1. Introduction

Similarity analysis is an important research area for DNA sequences. One of the most important bioinformatics research topics is DNA sequence similarity analysis [1–4]. DNA sequence analysis is the first step in identifying similar nucleotide sequences within a large genomic data repository, and is used for identifying many evolutionary or affinity relations and pathophysiological processes. A number of methods have recently been proposed to accurately and effectively determine the similarity in DNA sequences [1]. Alignment-based similarity analysis is one of the two main topics in this area, and many algorithms and tools have been developed for it [5–8]. Alignment-based methods are generally based on finding the optimal alignment result using search, gapping, and shifting operations. However, they have computational costs and are time-consuming.

These disadvantages have induced researchers to focus on different methods, and alignment-free methods have been proposed. Alignment-free methods convert DNA sequences into digitized vectors for numerical characterization; the similarity between these vectors can then be calculated. Graphical-based methods, information theory-based methods, graph-based methods, and word-frequency-based methods, etc., are used for digitization processing [9–16]. DNA sequence similarity analysis attempts to calculate the similarities between two or more sequences. Many similarity calculation methods can be applied to DNA similarity analysis, with varying levels of success. In this context, we propose a method based on the application of image texture analysis, which has a wide area of study in computer science, for use in DNA sequence similarity analysis. Texture, which is the pattern of information or arrangement of the structure found in an image, is an important feature of many image types. Generally, texture refers to the surface characteristic and appearance of an object defined by the shape, size, arrangement, density, and proportion of its elementary parts.

Due to the information contained in the texture, texture feature extraction is a key function in various image processing applications, remote sensing, and content-based image retrieval. Texture

* Corresponding author.
E-mail addresses: edelibas@cumhuriyet.edu.tr (E. Delibaş), ahmetarslan@selcuk.edu.tr (A. Arslan).

https://doi.org/10.1016/j.jmgm.2020.107603
1093-3263/

```
>J01859.1 Escherichia coli 16S ribosomal RNA, complete sequence
AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGT
AACAGGAAGAAGCTTGCTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATG
GAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAAGAGGGGGACCTTCG
GGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAACGGCTCACCTAGGCGACG
ATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGG
CAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTT
CGGGTTGTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTACCCG
CAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAAT
TACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAAC
TGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGT
AGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCG
TGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCC
TTGAGGCGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACT
CAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCT
TACCTGGTCTTGACATCCACGGAAGTTTTCAGAGATGAGAATGTGCCTTCGGGAACCGTGAGACAGGTGC
TGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTATCCT
TTGTTGCCAGCGGTCCGGCCGGGAACTCAAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGA
CGTCAAGTCATCATGGCCCTTACGACCAGGGCTACACACGTGCTACAATGGCGCATACAAAGAGAAGCGA
CCTCGCGAGAGCAAGCGGACCTCATAAAGTGCGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATG
AAGTCGGAATCGCTAGTAATCGTGGATCAGAATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACCG
CCCGTCACACCATGGGAGTGGGTTGCAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCACTT
TGTGATTCATGACTGGGGTGAAGTCGTAACAAGGTAACCGTAGGGGAACCTGCGGTTGGATCACCTCCTT
```

Fig. 1. 16S ribosomal DNA of *Escherichia coli* with FASTA Format.

Table 1
16S ribosomal DNA of 13 bacteria.

	Species	Accession Code	Length (bp)
1	*Bacillus maritimus*	KP317497	1515
2	*Bacillus wakoensis*	NR_040849	1524
3	*Bacillus australimaris*	NR_148787	1513
4	*Bacillus xiamenensis*	NR_148244	1513
5	*Escherichia coli*	J01859	1541
6	*Streptococcus himalayensis*	NR_156072	1509
7	*Streptococcus halotolerans*	NR_152063	1520
8	*Streptococcus tangierensis*	NR_134818	1520
9	*Streptococcus cameli*	NR_134817	1518
10	*Thermus amyloliquefaciens*	NR_136784	1514
11	*Thermus tengchongensis*	NR_132306	1523
12	*Thermus thermophilus*	NR_037066	1515
13	*Thermus filiformis*	NR_117152	1514

Table 2
Feature vectors obtained from the textures converted from the 16S ribosomal DNA of 13 bacteria.

	Skewness	Kurtosis	Energy	Entropy
Bacillus Australimaris	0,1596	−1,1015	0,0667	3,9518
Bacillus Maritimus	0,1027	−0,9176	0,0739	3,9087
Bacillus Wakoensis	0,1452	−1,1183	0,0640	4,0106
Bacillus Xiamenensis	0,1144	−0,9550	0,0724	3,9285
Escherichia coli	0,1267	−0,9394	0,0729	3,9164
Streptococcus Cameli	0,1324	−1,1364	0,0635	4,0209
Streptococcus Halotolerans	0,1301	−1,1444	0,0636	4,0182
Streptococcus Himalayensis	0,1425	−0,9433	0,0728	3,9240
Streptococcus Tangierensis	0,1627	−1,1002	0,0645	4,0103
Thermus Amyloliquef	0,1605	−1,1020	0,0645	4,0117
Thermus Filiformis	0,1382	−1,1553	0,0635	4,0188
Thermus Tengchongensis	0,1550	−1,0887	0,0647	4,0079
Thermus Thermophilus	0,1469	−1,1399	0,0633	4,0270

features can be extracted by several methods using statistical, structural, model-based and transformed information [17]. The statistical methods used for feature extraction of image textures represent the texture according to the non-deterministic properties that indirectly govern the distributions and relationships between the gray levels of the image. This technique was one of the first methods for machine vision [18]. Statistical methods can be used to analyze the spatial distribution of gray-level values by calculating local characteristics at each point in the image and obtaining a set of statistics from the distribution of the local characteristics. Histogram-based features, which are a first-order statistic, are calculated from the original image features and do not take into account the neighborhood relationships. The histogram-based approach to texture analysis is based on concentrations of intensity values for a whole image or a portion of an image that are shown as a histogram. Common features include the average, variance, energy, entropy, skewness, and kurtosis [19]. Our method applies the theory of texture analysis based on these moments to identify and calculate the features of a DNA sequence.

The histogram of a given image can be easily calculated. The shape of a histogram contains substantial information about the

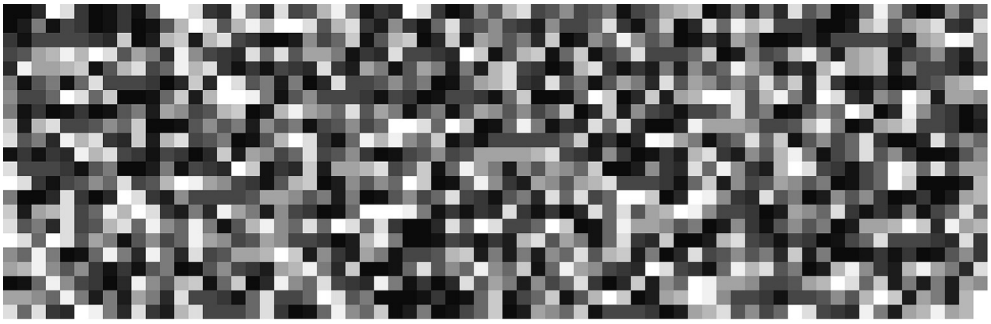

Fig. 2. Textures converted from the DNA sequences shown in Fig. 1.

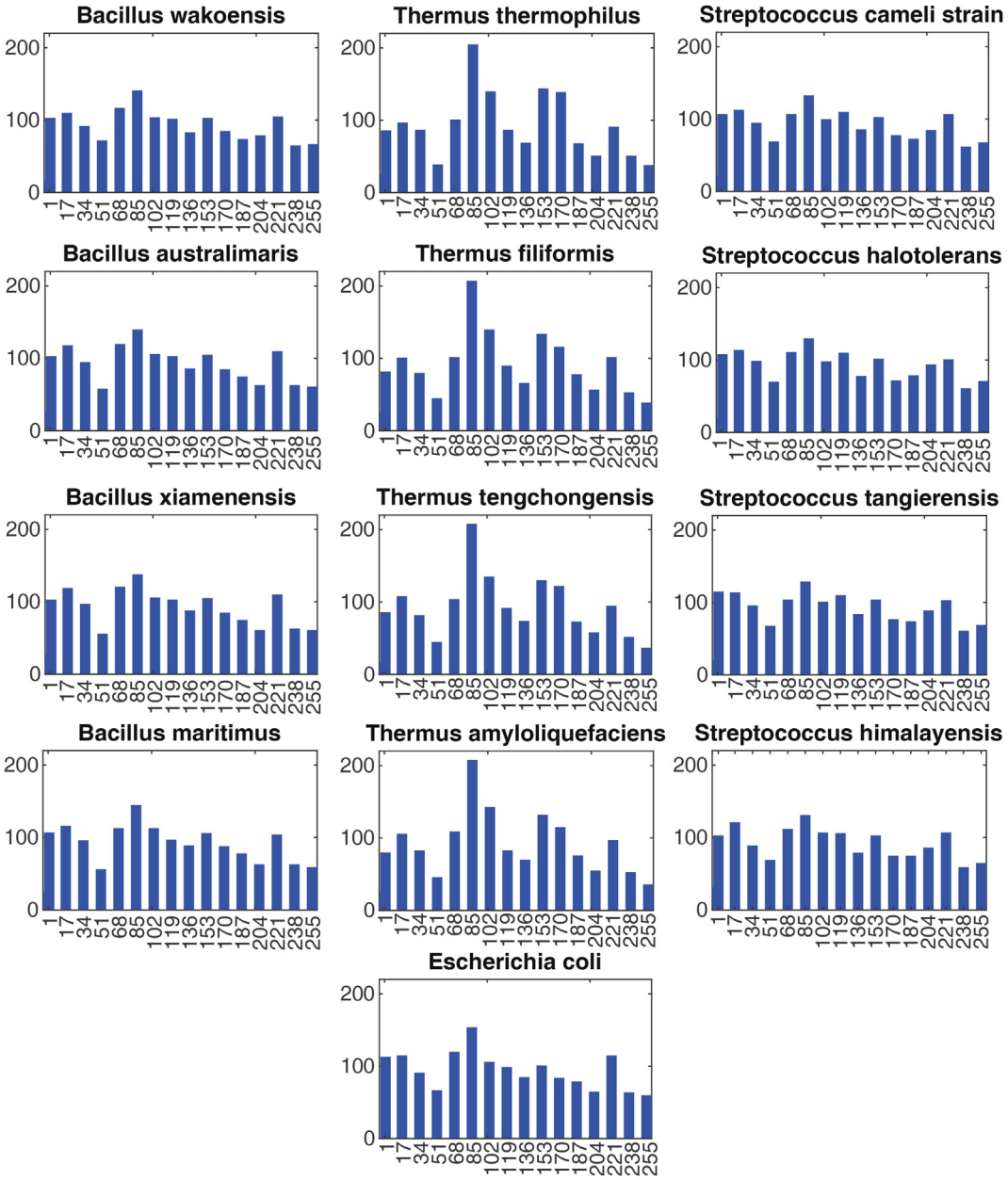

Fig. 3. Histograms of the textures converted from the DNA sequences presented in Table 1.

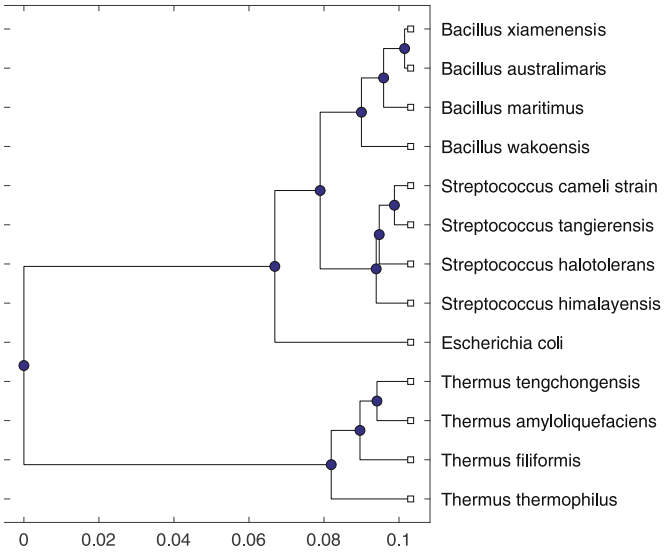

Fig. 4. Phylogenetic tree generated by the proposed method.

image. For example, a narrowly distributed histogram indicates a low-contrast image. A bimodal histogram generally indicates that the image contains an object of narrow intensity against a background of different intensity [19]. Texture can be characterized by the information contained in the histogram. In this way, histogram-based features are frequently used in the classification of images, image retrieval, and image segmentation [20]. In addition, texture analysis has been used in many biomedical applications such as for the detection of breast lesions [21], in vessel segmentation techniques [22], and in staging PET/CT images [23].

In this paper, unlike the study of Chen et al. [24], which used the Gray-Level Co-occurrence Matrix (GLCM) method, the histogram-based feature extraction method for image textures is applied to

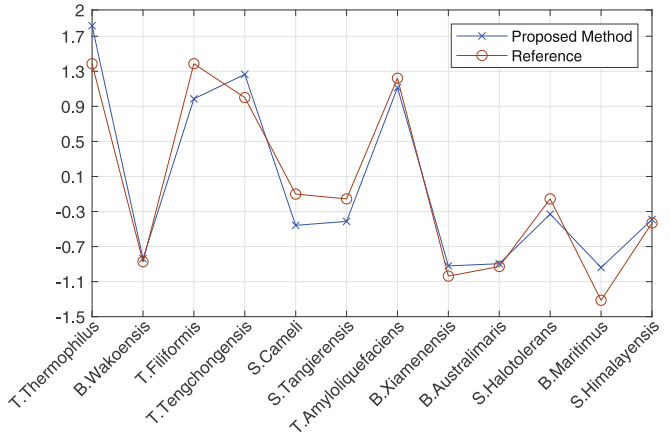

Fig. 6. The degree of similarity/dissimilarity of the other 12 bacteria and *Escherichia coli.*

DNA sequence analysis. As in alignment-free DNA sequence analysis methods, each sequence is subjected to vector digitization independently of one another. The analysis is then conducted using the similarity metrics.

2. Methods

2.1. Converting a DNA sequence to a digital image

Similar to pixels forming a picture, a DNA sequence consists of sub-units of characters: A (Adenine), G (Guanine), C (Cytosine), and T (Thymine). For a gray-level image, pixels are represented by numbers corresponding to the gray-level value of that pixel. For converting the DNA sequence to an image, numerical equivalents are assigned for the letters corresponding to the nucleotides. Histogram-based features, as mentioned above, are calculated from

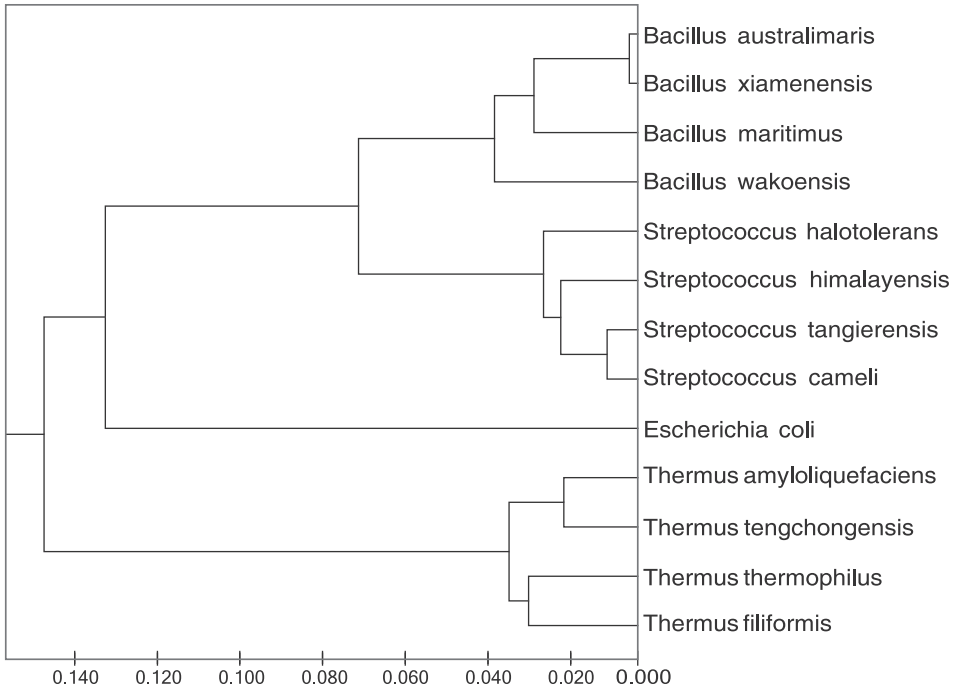

Fig. 5. Phylogenetic tree generated by MEGA7 based on ClustalW alignment and the UPGMA method.

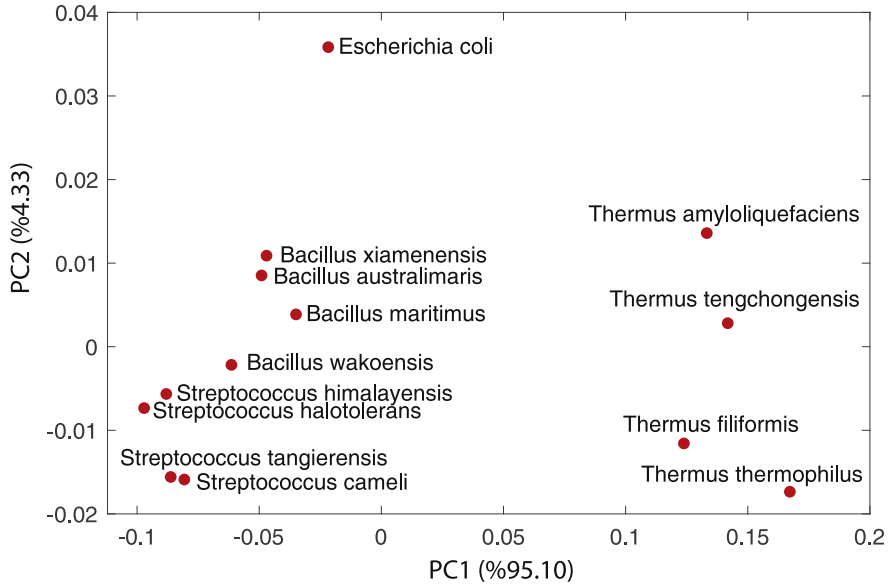

Fig. 7. The projection of the 4-dimensional vectors of 13 bacteria into 2D space consisting of two principal components PC1 and PC2.

Table 3
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

the original image features and do not take into account the neighborhood relationships. For this reason, to further strengthen the characterization, we use nucleotides in binary groups. Thus, we also embed data about the base transitions as a neighborhood into the original image features. When assigning the values to the letters, we determine the numerical response for the 16 cases, which are the binary combinations of the 4 letters. Valued binary groups of nucleotides are obtained by shifting one base. These dinucleotides are as follows:

$$\alpha = \{AA, AG, AC, AT, GA, GG, GC, GT, CA, CG, CC, CT, TA, TG, TC, TT\}$$

Each of the dinucleotides is delivered equally spaced between the gray level value of 1–255, respectively.

DNA sequences can be accessed from genetic databases with an accession code. When one exports the DNA sequence into the

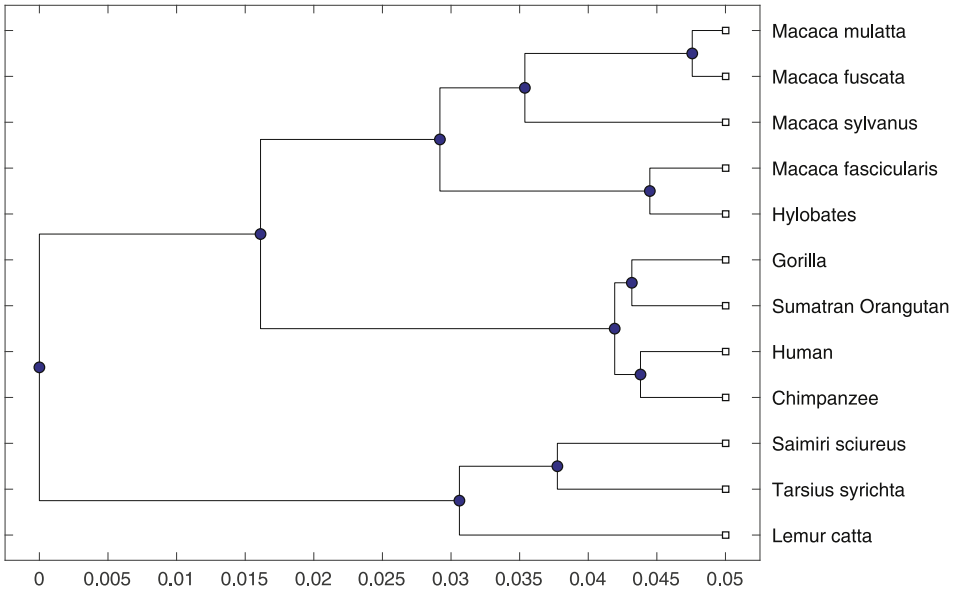

Fig. 8. Phylogenetic tree generated by the proposed method.

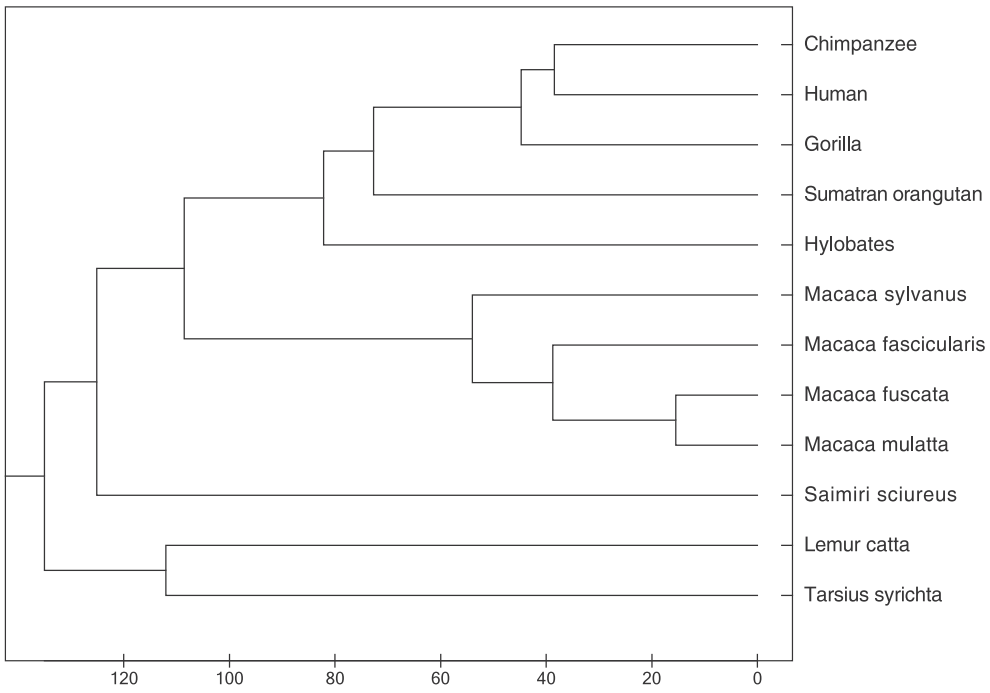

Fig. 9. Phylogenetic tree generated by MEGA7 based on ClustalW alignment and the UPGMA method.

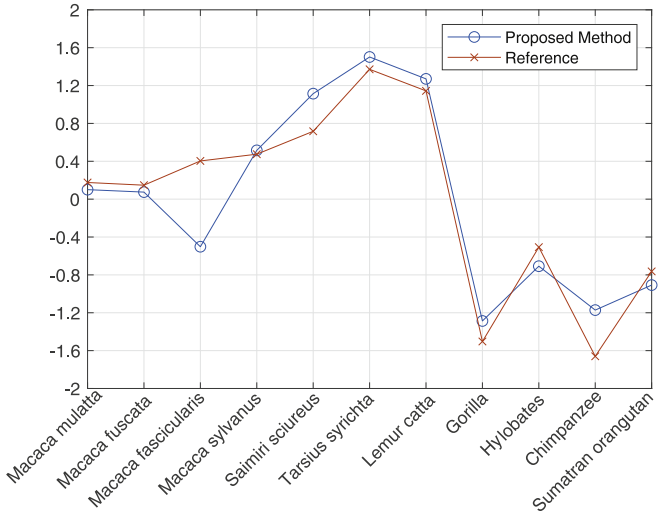

Fig. 10. The degree of similarity/dissimilarity of the other 11 species and Human.

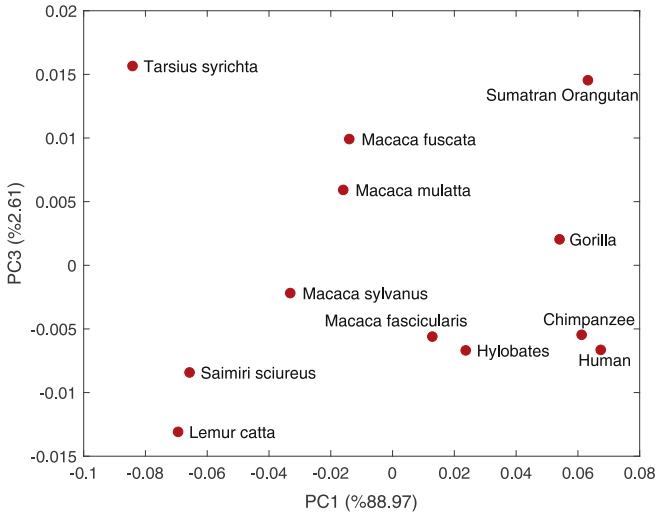

Fig. 11. The projection of the 4-dimensional vectors of 12 species into 2D space consisting of two principal components PC1 and PC3.

FASTA format, it is arranged to have 70 bases in a row. The conversion to the image is created according to the sequence in this format. In this way, the image is obtained with a line length of 70 pixels in width and with a height varying according to the length of the sequence. This format can be seen in Fig. 1, which depicts a sample 16S ribosomal DNA sequence. In this way, a gray-level image is obtained, and all the sequences to be analyzed are transformed.

2.2. Feature extraction based on histogram analysis

Assume the image is a function $f(x,y)$ of two space variables x and y, $x = 0, 1, \ldots, N-1$ and $y = 0, 1, \ldots, M-1$. The function $f(x,y)$ can take discrete values of $i = 0, 1, \ldots, G-1$, where G is the total number of intensity levels in the image. The intensity-level histogram is a function showing (for each intensity level) the number of pixels in the whole image, which have the intensity:

$$h(i) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \delta(f(x,y), i) \tag{1}$$

where $\delta(j,i)$ is the Kronecker delta function.

$$\delta(j,i) = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \tag{2}$$

Table 4
The 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

The histogram of intensity levels is a brief and simple summary of the statistical information available in the image. The calculation of the gray-level histogram involves single pixels. Therefore, the histogram contains the first-order statistical information about the image or its fragment. The division of the $h(i)$ values by the total number of pixels in the image yields an approximate probability density of occurrence of the intensity levels [25].

$$p(i) = h(i)/NM, \quad i = 0, 1, \ldots, G-1 \tag{3}$$

First-order texture analysis (or histogram analysis) extracts pixel intensity values within a gray-level image [26]. Different parameters are used to define the first-order statistic features of the image over these values that calculate the histogram. These metrics, also called central moments [27], are also values that will characterize the DNA sequence [28,29]. The equations are given as Equations (4)–(9) [25].

$$\mu = \sum_{i=0}^{G-1} ip(i) \tag{4}$$

$$\sigma^2 = \sum_{i=0}^{G-1} (i-\mu)^2 p(i) \tag{5}$$

$$\mu_3 = \sigma^{-3} \sum_{i=0}^{G-1} (i-\mu)^3 p(i) \tag{6}$$

$$\mu_4 = \sigma^{-4} \sum_{i=0}^{G-1} (i-\mu)^4 p(i) - 3 \tag{7}$$

$$E = \sum_{i=0}^{G-1} [p(i)]^2 \tag{8}$$

$$H = -\sum_{i=0}^{G-1} p(i) log_2[p(i)] \tag{9}$$

The mean value given in (4) refers to the average intensity level of the image or texture. The variance (5) also describes the variation of the density around the mean. The skewness value (6) is zero if the histogram is symmetrical around the mean, otherwise it may be positive or negative depending on whether it is less than or greater than the mean. Accordingly, μ3 is an indicator of symmetry. The kurtosis (7) is a measure of flatness of the histogram. The component '3' inserted in (7) normalizes μ4 to zero for a Gaussian-shaped histogram. The energy (8) is a measure of histogram uniformity. Finally, the entropy (9) of an image is an estimation of randomness, and is frequently used to measure its texture. Entropy can be thought of as a measurement of the sharpness of the histogram peaks, which is directly related to more defined structural information [30].

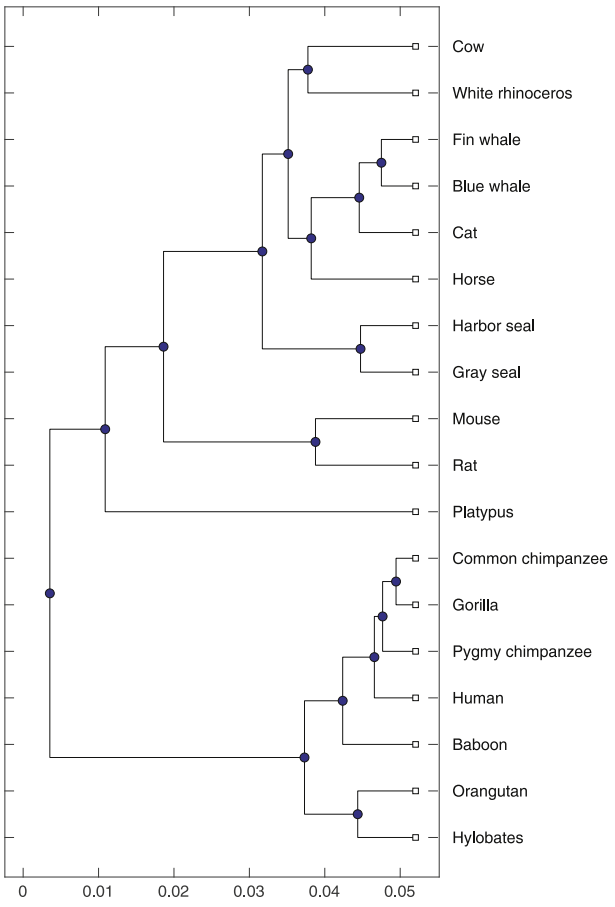

Fig. 12. Phylogenetic tree generated by the proposed method.

We design the feature vector with the metrics, which contain information about the shape of the histogram. Therefore, the DNA sequences are transformed into 4-dimensional vectors by the following metrics that characterize the features of the DNA sequences, which is then converted to an image. The vector V is given in (10):

$$V = [\mu_3, \mu_4, E, H] \tag{10}$$

2.3. Construction and comparison of the phylogenetic tree

We evaluated the similarity measurements of the proposed method using the "MATLAB Statistics and Machine Learning Toolbox" and "MATLAB Bioinformatics Toolbox" to perform clustering. Using the feature vectors as in Table 2, we generated the phylogenetic tree using the functions "pdist" with default

parameters (Euclidean), and "seqlinkage" with "average" parameter to generate the dendrograms in MATLAB R2018b. We then calculated the similarities between the DNA sequences. As a reference result for comparison to the phylogenetic tree, we used MEGA7 [31], which is the Molecular Evolutionary Genetics Analysis software.

2.4. Similarity calculation

We obtain a characterization vector V in the 4-dimensional linear space, followed by a comparison between histograms of the textures generated by sequences with these vectors. Similarities between these vectors can be calculated by applying the Euclidean distance (11) between their end points:

$$E = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{11}$$

3. Results

3.1. Data description

We apply our method on three DNA datasets, and compare the results with the reference trees genereated by MEGA software. We have chosen those datasets of different number and lengths of sequences. Firstly, we created a data set ourselves and compared it with the reference tree. We also used two data sets used by previous researchers to evaluate the results.

3.2. Implementation

3.2.1. 16S ribosomal DNA of bacteria

We choose the 16S ribosomal DNA of 13 bacteria to test our method. Bacteria were selected randomly from three distinct groups, along with a single bacterium to test sequences that are very similar, along with well-separated sequences. All sequences were selected from the NCBI database and are listed in Table 1. The sequence lengths are between 1509 and 1541 bases.

We applied our method to the 16S ribosomal DNA sequences presented in Table 1. The texture converted from the sequence in Fig. 1 is presented in Fig. 2 as a sample.

We calculated the histograms for all of the textures converted from the DNA sequences given in Table 1. The resulting histograms are shown in Fig. 3. The topological structures of groups in DNA sequences can be roughly seen from the similarities of histograms.

Table 2 presents the feature vectors obtained from calculating the histograms of the textures converted from the 16S ribosomal DNA textures of 13 bacteria.

We calculated the similarities between bacteria from feature vectors presented in Table 2. The phylogenetic tree based on this

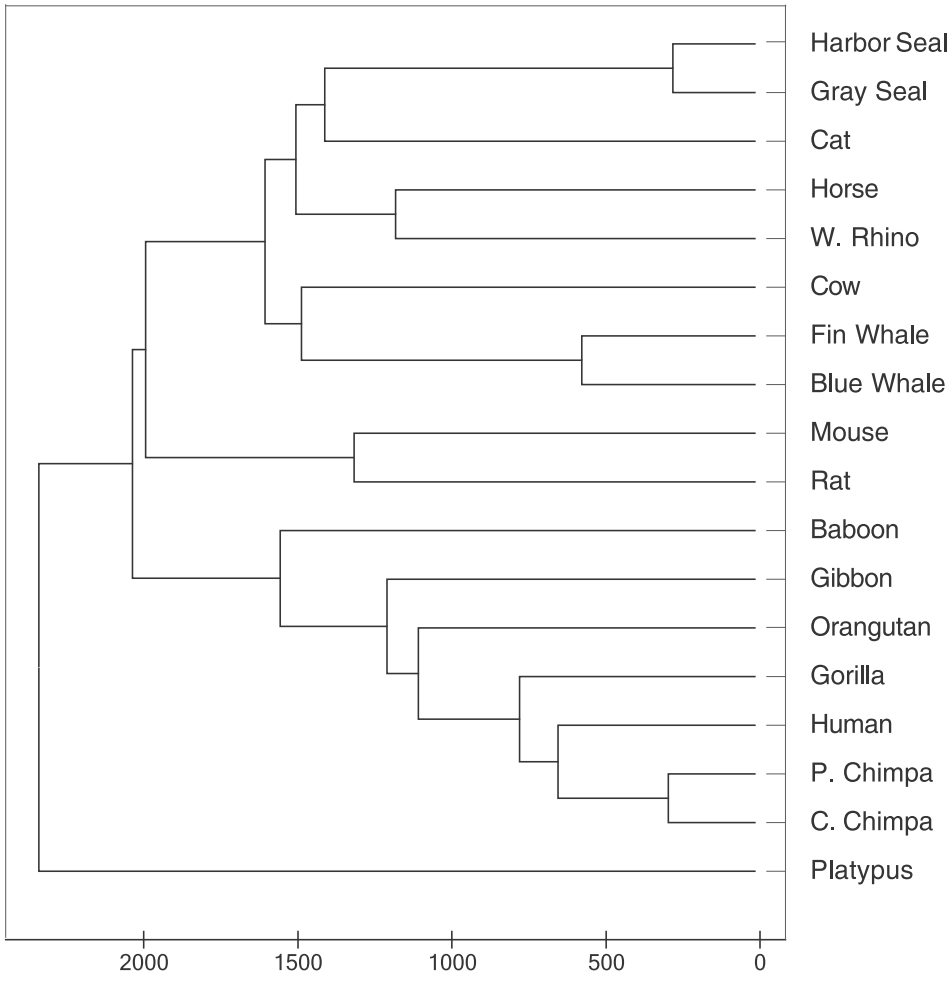

Fig. 13. Phylogenetic tree generated by MEGA7 based on ClustalW alignment and the UPGMA method.

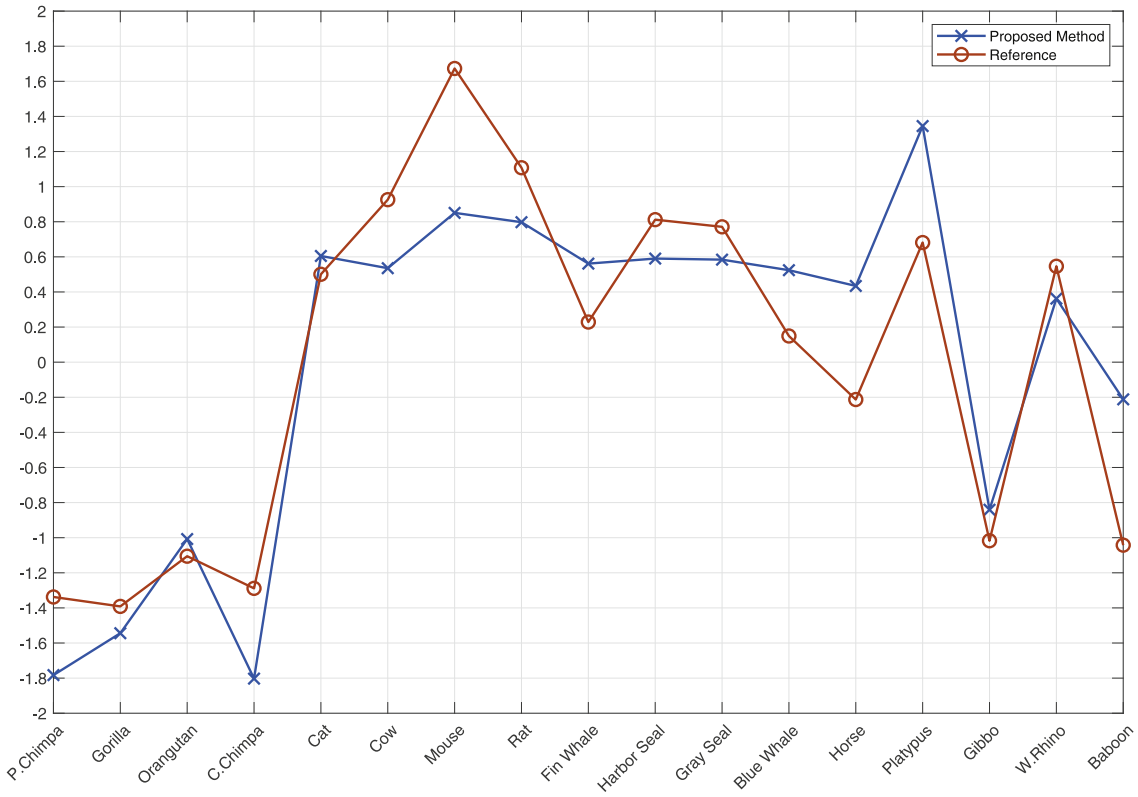

Fig. 14. The degree of similarity/dissimilarity of the other 12 species and Human.

similarity calculation is presented in Fig. 4. The DNA sequences shown in Table 1 were analyzed in MEGA7 by the alignment-based method ClustalW. The UPGMA method was used to construct the phylogenetic trees, shown in Fig. 5.

Considering the accuracy, the dendrogram tree generated with the proposed method has same topology and is consistent with reference tree. We can also see the overall agreement about consonance of similarity/dissimilarity matrix of proposed method and the reference tree generated by MEGA7 qualitatively. To reveal and visualize this, we denote the degree of similarity/dissimilarity between *Escherichia coli* and other species by distance measurement plots in Fig. 6. We can see that the curvilinear tendency of these two curves is almost the same, indicating the general agreement between the similarities/dissimilarities obtained by the proposed method and the reference.

To verify the effectiveness of the high-dimensional vectors in Table 2 that we produced with our method, we perform principal component analysis (PCA) on the 4 parameters in the vector. In Fig. 7, the projection of 13 vectors into a 2D property space consisting of two main components PC1, PC2. In Fig. 7 we can see that the results are overall in agreement with the results above. It is also worth noting that the first two PCAs contain 99.43% of the total inertia of the 4-dimensional space vector. This rough projection also confirms that the mathematical identifier effectively characterizes the DNA sequence structure.

3.2.2. NADH dehydrogenase subunit 4 genes

Another data set we use is the NADH dehydrogenase subunit 4 genes of 12 species of 4 different groups of primates. The dataset consists of 4 species of old-world monkeys, one specie of new-world monkeys, two species of prosimians and five species of hominoids. All the sequences are obtained from NCBI database and listed in Table 3, whose lengths are between 893 and 896 base pairs. They were previously reported and used in their methods by Hayasaka et al. [32], subsequently used by Zhang [33,34], Qi et al. [14] and Chen et al. [24].

We also calculated histograms and obtained the feature vectors for all textures converted from these DNA sequences presented in Table 3. After calculating the similarities between the species from feature vectors, we generated the phylogenetic tree based on this similarity. In Fig. 8, we presented the phylogenetic tree generated by our method and we also presented the reference phylogenetic tree obtained by MEGA7 based on ClustalW aligment and UPGMA method in Fig. 9.

We can also see the overall agreement about consonance of similarity/dissimilarity matrix of proposed method and reference tree generated by MEGA7 qualitatively. To reveal and visualize this, we also denoted the degree of similarity/dissimilarity between *Human* and other species by distance measurement plots in Fig. 10. We can see that the curvilinear tendency of these two curves is similar, indicating the general agreement between the similarities/dissimilarities obtained by the proposed method and the reference. In Fig. 11 we presented the projection of 12 feature vectors into 2D property space consisting of two main components PC1 and PC3. In the figure we can also see that the results are overall in agreement with the results above. Two PCAs contain 91.58% of the total inertia of 4-dimensional vector of this data set.

3.2.3. Whole mitochondrial genomes of 18 eutherian mammals

The whole mitochondrial genomes include abundant genetic information, that used frequently in recent years, of 18 eutherian mammals [35]. All of the sequences are obtained from NCBI database and listed in Table 4, whose lengths are between 16295 bases and 17019 bases.

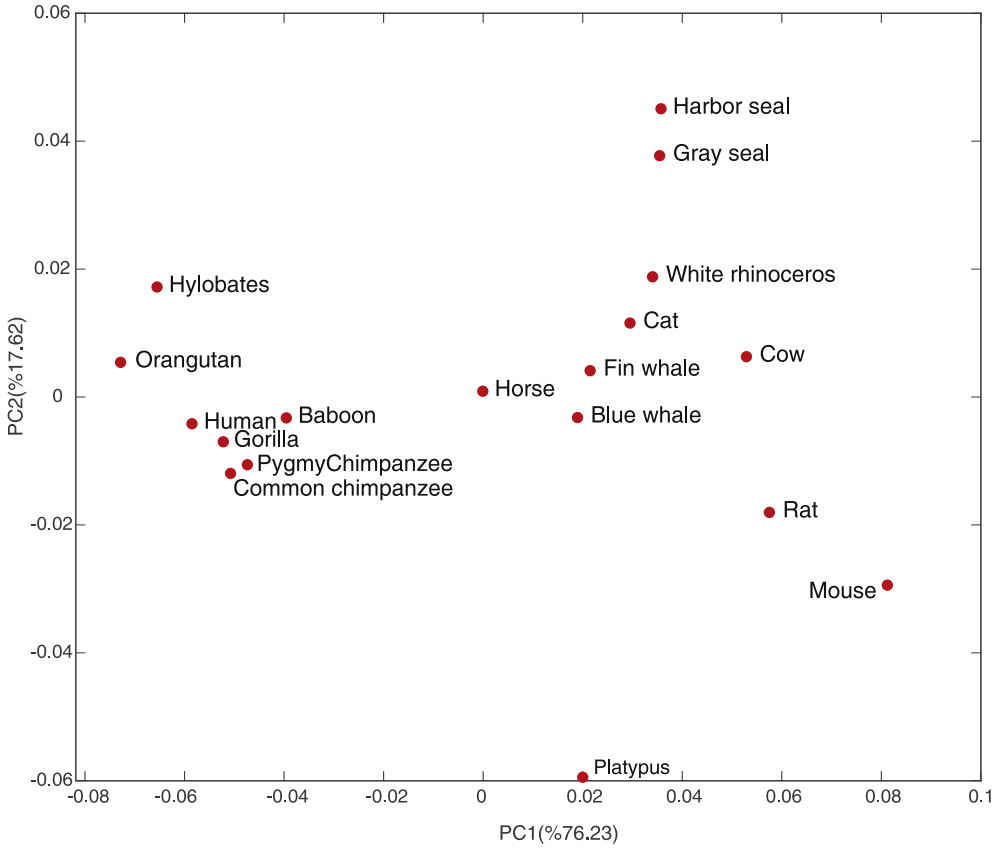

Fig. 15. The projection of the 4-dimensional vectors of 18 eutherian mammals into 2D space consisting of two principal components PC1 and PC2.

We also calculated the histograms and obtained the feature vectors for all textures converted from these DNA sequences presented in Table 4. After calculating the similarities between species from feature vectors, we generated the phylogenetic tree based on this similarity. In Fig. 12, we presented the phylogenetic tree generated by our method and we also presented the reference phylogenetic tree obtained by MEGA7 based on ClustalW aligment and UPGMA method in Fig. 13.

We can also see the overall agreement about consonance of similarity/dissimilarity matrix of proposed method and reference tree generated by MEGA7 qualitatively for this data set. To reveal and visualize this, we denoted the degree of similarity/dissimilarity between *Human* and the other 17 species by distance measurement plots in Fig. 14. We can also see that the curvilinear tendency of these two curves is similar, indicating the general agreement between the similarities/dissimilarities obtained by the proposed method and the reference. In Fig. 15 we presented the projection of 18 feature vectors into 2D property space consisting of two main components PC1 and PC2. In the figure we can also see that the results are overall in agreement with the results above. Two PCAs contain 93.85% of the total inertia of 4-dimensional vector of this data set.

4. Conclusion

In this paper, a different approach for DNA sequence similarity analysis performed on the basis of similarity calculations is introduced. In computer science, various similarity calculations are studied in processes such as classification, recognition, and segmentation. Texture analysis methods, which are a subset of digital image processing methods, are used with the assumption that these calculations can be adapted to alignment-free DNA sequence similarity analysis. Similarity calculations are made between the DNA sequences, which are converted into images via histogram-based texture analysis based on calculations from a single pixel. In the texture analysis, regular or repeated patterns form the characteristic of textures. In this study, we showed that the pattern of nucleotides can characterize DNA sequences. We obtained texture features for DNA sequences, and a similarity matrix was computed. The phylogenetic relationships revealed by our method showed that our tree was similar to the result of MEGA, which is based on sequence alignment. These findings show that texture analysis metrics can also be used to characterize DNA sequences.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmgm.2020.107603.

References

[1] X. Jin, Q. Jiang, Y. Chen, S.J. Lee, R. Nie, S. Yao, et al., Similarity/dissimilarity calculation methods of dna sequences: a survey, J. Mol. Graph. Model. 76 (Supplement C) (2017) 342–355.
[2] O. Bonham-Carter, J. Steele, D. Bastola, Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis, Briefings Bioinf. 15 (6) (2014) 890–905.
[3] S. Vinga, J. Almeida, Alignment-free sequence comparison - a review, Bioinformatics 19 (4) (2003) 513–523.
[4] A. Zielezinski, S. Vinga, J. Almeida, W.M. Karlowski, Alignment-free sequence comparison: benefits, applications, and tools, Genome Biol. 18 (1) (2017) 186.
[5] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.
[6] S.B. Needleman, C.D. Wunsch, A general method applicable to search for similarities in amino acid sequence of 2 proteins, J. Mol. Biol. 48 (3) (1970)

443.
[7] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. U. S. A. 85 (8) (1988) 2444–2448.
[8] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1) (1981) 195–197.
[9] M. Rinku, A. Neeru, A graph theoretic model for prediction of reticulation events and phylogenetic networks for dna sequences, Egypt. J. Basic Appl. Sci. 3 (3) (2016) 263–271.
[10] Y.H. Yao, S.J. Yan, H.M. Xu, J.N. Han, X.Y. Nan, P.A. He, et al., Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation, Evol. Bioinf. Online 10 (2014) 87–96.
[11] B. Liao, Q.L. Xiang, L.J. Cai, Z. Cao, A new graphical coding of dna sequence and its similarity calculation, Phys. Stat. Mech. Appl. 392 (19) (2013) 4663–4667.
[12] N. Jafarzadeh, A. Iranmanesh, C-curve: a novel 3d graphical representation of dna sequence based on codons, Math. Biosci. 241 (2) (2013) 217–224.
[13] I. Soares, A. Goios, A. Amorim, Sequence comparison alignment-free approach based on suffix tree and l-words frequency, Sci. World J. 2012 (2012), https://doi.org/10.1100/2012/450124, 450124-450124.
[14] X.Q. Qi, Q. Wu, Y.S. Zhang, E. Fuller, C.Q. Zhang, A novel model for dna sequence similarity analysis based on graph theory, Evol. Bioinf. Online 7 (2011) 149–158.
[15] J.F. Yu, J.H. Wang, X. Sun, Analysis of similarities/dissimilarities of dna sequences based on a novel graphical representation, Match Commun. Math. Comput. Chem. 63 (2) (2010) 493–512.
[16] J.F. Yu, X. Sun, J.H. Wang, Tn curve: a novel 3d graphical representation of dna sequence based on trinucleotides and its applications, J. Theor. Biol. 261 (3) (2009) 459–468.
[17] T.A. Pham, Optimization of Texture Feature Extraction Algorithm, Master of science, 2010.
[18] M. Tuceryan, A.K. Jain, Texture Analysis, 1998, pp. 207–248.
[19] G.N. Srinivasan, S. G, Statistical texture analysis, Int. J. Comput. Info. Eng. 2 (12) (2008) 4268–4273.
[20] T. Mapayi, S. Viriri, J.R. Tapamo, Adaptive thresholding technique for retinal vessel segmentation based on glcm-energy information, Comput. Math. Methods Med. (2015).
[21] W. Gomez, W.C.A. Pereira, A.F.C. Infantosi, Analysis of co-occurrence texture statistics as a function of gray-level quantization for classifying breast ultrasound, IEEE Trans. Med. Imag. 31 (10) (2012) 1889–1899.
[22] M.M. Fraz, S.A. Barman, P. Remagnino, A. Hoppe, A. Basit, B. Uyyanonvara, et al., An approach to localize the retinal blood vessels using bit planes and centerline detection, Comput. Methods Progr. Biomed. 108 (2) (2012) 600–616.
[23] T. Win, K.A. Miles, S.M. Janes, B. Ganeshan, M. Shastry, R. Endozo, et al., Tumor heterogeneity and permeability as measured on the ct component of pet/ct predict survival in patients with non–small cell lung cancer, Clin. Canc. Res. 19 (13) (2013) 3591.
[24] W.Y. Chen, B. Liao, W.W. Li, Use of image texture analysis to find dna sequence similarities, J. Theor. Biol. 455 (2018) 1–6.
[25] A. Materka, M. Strzelecki, Texture Analysis Methods – a Review. Report, Technical University of Lodz, Institute of Electronics, 1998.
[26] S. Alobaidli, S. McQuaid, C. South, V. Prakash, P. Evans, A. Nisbet, The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment planning, British J. Radiol. 87 (1042) (2014) 20140369.
[27] A. Papoulis, Probability, Random Variables and Stochastic Processes, McGraw-Hill, 1965.
[28] M. Levine, Vision in Man and Machine, McGraw-Hill, 1985.
[29] W. Pratt, Digital Image Processing, Wiley, 1991.
[30] E. Vazquez-Fernandez, A. Dacal-Nieto, F. Martín-Rodríguez, S. Torres-Guijarro, Entropy of Gabor Filtering for Image Quality Assessment, 2010.
[31] S. Kumar, G. Stecher, K. Tamura, Mega7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, Mol. Biol. Evol. 33 (7) (2016) 1870–1874.
[32] K. Hayasaka, T. Gojobori, S. Horai, Molecular phylogeny and evolution of primate mitochondrial-dna, Mol. Biol. Evol. 5 (6) (1988) 626–644.
[33] Y.S. Zhang, A simple method to construct the similarity matrices of dna sequences, Match Commun. Math. Comput. Chem. 60 (2) (2008) 313–324.
[34] Y.S. Zhang, W. Chen, New invariant of dna sequences, Match Commun. Math. Comput. Chem. 58 (1) (2007) 197–208.
[35] X. Jin, R. Nie, Dongming Zhou, Shaowen Yao, Yanyan Chen, Jiefu Yu, Quan Wang, A novel dna sequence similarity calculation based on simplified pulse-coupled neural network and huffman coding, Phys. Stat. Mech. Appl. 461 (2016) 325–338.