Journal of Molecular Graphics and Modelling 100 (2020) 107693

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

A novel alignment-free DNA sequence similarity analysis approach based on top-*k* *n*-gram match-up

Emre Delibaş [a, *], Ahmet Arslan [b], Abdulkadir Şeker [a], Banu Diri [c]

[a] Department of Computer Engineering, Faculty of Engineering, Sivas Cumhuriyet University, 58140, Sivas, Turkey
[b] Department of Computer Engineering, Faculty of Engineering, Selçuk University, 42250, Konya, Turkey
[c] Department of Computer Engineering, Faculty of Electrical and Electronics, Yıldız Technical University, 34349, İstanbul, Turkey

ARTICLE INFO

Article history:
Received 12 April 2020
Received in revised form
15 June 2020
Accepted 6 July 2020
Available online 7 August 2020

Keywords:
DNA sequence Similarity
Top-*k* *n*-gram
Alignment-free comparison

ABSTRACT

DNA sequence similarity analysis is an essential task 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. As an alternative to alignment-based sequence comparison methods, which result in high computational cost, alignment-free methods have emerged that calculate similarity by digitizing the sequence in a different space. In this paper, we proposed an alignment-free DNA sequence similarity analysis method based on top-*k* *n*-gram matches, with the prediction that common repeating DNA subsections indicate high similarity between DNA sequences. In our method, we determined DNA sequence similarities by measuring similarity among feature vectors created according to top-*k* *n*-gram match-up scores without the use of similarity functions. We applied the similarity calculation for three different DNA data sets of different lengths. The phylogenetic relationships revealed by our method show that our trees coincide almost completely with the results of the MEGA software, which is based on sequence alignment. Our findings show that a certain number of frequently recurring common sequence patterns have the power to characterize DNA sequences.

1. Introduction

As a result of rapid technological developments in recent decades, a large amount of biological data has become available, and research on most analytical methods has sought to extract information from this large quantity of data. DNA sequence analysis has the potential to play a key role in these studies. DNA sequence analysis is the first step in identifying similar nucleotide sequences, many evolution or affinity relationships, pathophysiological processes, etc., within the large genomic data repositories. Similarity analysis of DNA sequences is a process of inference that compares unknown sequences against known sequences to obtain the functions of the unknown ones [1]. Computational methods for DNA sequence similarity analysis provide the most successful solutions for difficult biological analysis.

A number of methods have been proposed for effectively and accurately identifying similarity in DNA sequences 2. The methods used in similarity analysis of DNA sequences are discussed in two main groups. The first is the alignment-based similarity analysis [3–6] which has been as the traditional method of such analyses, and the second is the alignment-free similarity analysis proposed to reduce computational complexity. The main objective of the alignment-free methods proposed for this purpose is to convert DNA sequences into digitized vectors for numerical characterization. The similarity between these vectors can then be calculated. Information theory-based methods, 2-D/3-D graphical representation-based methods, graph-based methods, chemical properties of nucleotide base compositions-based methods, etc., are used for digitization processing [7–14].

Methods using word-based measurement are also among the most widely used alignment-free methods [15–17]. Among word-based methods, the *n*-gram-based approach was first proposed by Blaisdell [18]. *N*-gram-based methods can be used quickly and easily for phylogenetic analysis of genetic sequences, without requiring evolutionary models [19,20]. Numerous *n*-gram-based methods have been proposed and used for sequence comparison

* Corresponding author.
E-mail addresses: edelibas@cumhuriyet.edu.tr (E. Delibaş), ahmetarslan@selcuk.edu.tr (A. Arslan), aseker@cumhuriyet.edu.tr (A. Şeker), banu@ce.yildiz.edu.tr (B. Diri).

https://doi.org/10.1016/j.jmgm.2020.107693
1093-3263/

and phylogenetic analysis [16,21–25]. In addition, several approaches have been proposed for finding the optimal range of n [26–28]. Some methods directly use the *n*-gram frequency by measuring the distance between DNA sequences using similarity functions such as Jensen–Shannon divergence and Kullback–Leibler divergence [29–31].

In order to reduce the computational cost in various text operations using the *n*-gram method, some parts of the *n*-grams are ignored. While doing this, the region which is called top-*k* *n*-gram and gets the highest score according to *n*-gram frequencies is determined and the dominant effect of this region on the similarity calculation is used [32–34]. In this paper, we use the top-*k* *n*-gram match-up numbers of DNA sequences as pairwise evaluations to obtain feature vectors. A hybrid approach has emerged on the background of the method, inspired by previous studies, using direct *n*-gram frequencies and similarity calculation operations from the feature vector. It seeks to identify and use the most effective base groups to reveal DNA sequence similarity by using top-*k* *n*-gram. This significantly reduces computational costs.

2. Methods

2.1. N-gram analysis

N-grams are sequences of *n* words. In a biological context, it may be an *n*-gram, *n* amino acid or nucleotide sequence. For example, the "AGCTTAGCG" sequence has two counts of 2-grams of AG and GC, and one count of 2-grams of CT, TT, TA and CG. The formal definition of *n*-grams is given below:

Given a sequence of *N* words $S = (s_1, s_2, \ldots, s_N)$ over the vocabulary A, and a positive integer *n*, an *n*-gram of the sequence *S* is any subsequence $(s_i, s_{i+1}, \ldots, s_{i+n-1})$ of *n* consecutive words [35]. There are N-n+1 such *n*-grams in *S*. For a vocabulary *A* with $|A|$ distinct words, there are $|A|^n$ possible unique *n*-grams [19].

2.2. Top-k n-gram

In natural language processing (NLP) studies where *n*-gram is used, *n*-gram string synopses are employed to handle string predicates. An *n*-gram synopsis summarizes the vocabulary by "omitting" certain *n*-grams. Thus, a synopsis represents a subset of all possible *n*-grams occurring in the data. A simple strategy is to select random *n*-gram samples from the data. Another approach is to construct a top-*k* *n*-gram synopsis [33].

The top-*k* *n*-gram synopsis of document corpus S is an *n*-gram-to-count map, which consists of the set of *n*-grams in S with the highest counts. The construction method for top-*k* *n*-gram synopsis φ_{topk} is very straightforward:

- Find all unique *n*-grams and their counts in corpus S;
- Sort *n*-grams by counts;
- Insert the *n*-grams with the highest count into φ_{topk}, until the total size of the φ_{topk} exceeds the number (or percentage) of *k* of document corpus S [36].

2.3. Feature extraction based on top-k n-gram

A DNA sequence consists of a 4 letter alphabet: A (Adenine), G (Guanine), C (Cytosine), and T (Thymine). The *n*-gram, an NLP application, will be implemented on DNA sequences that we consider as a text consisting of these 4 letters. We use a top-k strategy to identify *n*-grams that will define DNA, because repetitive regions are distinctive features in the characterization of a DNA. Frequently repeating, high-frequency text fragments are valuable for characterizing DNA. With this approach, the use of frequently repeating high-frequency *n*-grams, instead of all *n*-grams in the sequence, will greatly alleviate the burden of using all *n*-grams and will enable us to use the essence of the DNA sequence in characterization. If *S* is a set of *n*-grams in the DNA sequence, $S = (s_1, s_2, \ldots, s_N)$; φ_{topk} is the first *k*-percentage in the ordered list, which decreases with respect to the frequencies of *n*-grams in the *S* corpus. If the sets of the top-*k* *n*-grams of the DNA sequences to be compared are $\varphi_{topk(i)}$, the feature vectors of these sequences are calculated as follows:

$$If\ \varphi_{topk(i,j)} \in \varphi_{topk(l)} \quad , \quad inc(m_{i,l})$$

where *i*, *l* = (1, 2, …, *d*), *d* is the number of DNA sequences to be compared; *j* = (1,2, …,*t*), *t* = N * *k*-percent; *m* is match-up score. The feature vectors $V_{(i)}$ that will occur for d = 5 are given in Table 1.

Unlike some methods that directly use the *n*-gram frequency by measuring the distance between DNA sequences using similarity functions, we used the match-up scores as a feature vector dimensions. We then calculated the similarities between the DNA sequences using distance measurement metrics from normalized values of these vectors. For instance, if AGCTCTACCG and AAAGCTTACTG are compared, 2-grams in descending order according to their frequency will be as follows:

$$S_1 = \{CT:2, AG:1, GC:1, TC:1, TA:1, AC:1, CC:1, CG:1\}$$
$$S_2 = \{AA:2, CT:2, AG:1, GC:1, TT:1, TA:1, AC:1, TG:1\}$$

For top25 2-gram, CT and AG are elements corresponding to 25 percentage of S_1 and form the $\varphi_{top25(1)}$ set. It is checked whether

Table 1
Feature vectors (V_i) of DNA sequences with match-up scores.

Sequences	1	2	3	4	5
1	1	$m_{1,2}$	$m_{1,3}$	$m_{1,4}$	$m_{1,5}$
2	$m_{2,1}$	1	$m_{2,3}$	$m_{2,4}$	$m_{2,5}$
3	$m_{3,1}$	$m_{3,2}$	1	$m_{3,4}$	$m_{3,5}$
4	$m_{4,1}$	$m_{4,2}$	$m_{4,3}$	1	$m_{4,5}$
5	$m_{5,1}$	$m_{5,2}$	$m_{5,3}$	$m_{5,4}$	1

Table 2
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

Table 3
Top-*k* *n*-gram accuracy and running time results for NADH dehydrogenase subunit 4 genes of 12 species.

n-gram	top-k (percent)	Accuracy	Comput. Time
9	7	100	0.1878
10	8	100	0.2023
13	8	100	0.2040
10	9	100	0.2143
13	9	100	0.2149

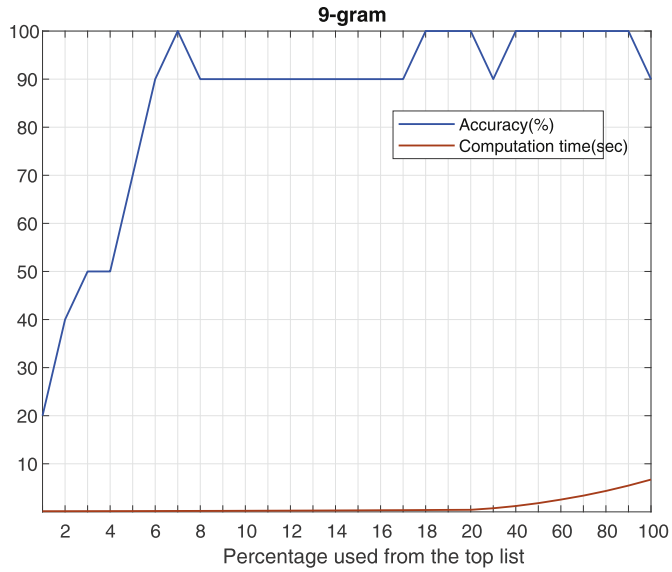

Fig. 1. Accuracy-computation time graph of 9-gram that produces the best results.

these elements are in $\varphi_{top25(2)}$ respectively. Accordingly, "CT" is the only matching element, and it is inserted into the relevant location in the vector V_1 as $m_{1,2} = 1$.

2.4. Normalizing the feature vectors

It is necessary to normalize the d-dimensional feature vectors $V_{(i)}$ obtained with match-up scores. Min-max normalization was used for 0–1 normalization. The equation for the min-max normalization is given in (1).

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

2.5. Similarity calculation

We obtain a characterization vector V in the d-dimensional linear space, followed by a comparison between sequences with these vectors. In calculating the similarity between DNA sequences, the distance calculation is the basis of the analysis and is an important step to represent the results. Euclidean distances is one of the most often used calculation methods [2]. Similarities between the characterization vectors we calculated by applying the Euclidean distance (2) between their end points:

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

2.6. Construction and comparison of the phylogenetic trees

We evaluated the similarity measurements of the proposed method using the "MATLAB Statistics and Machine Learning Toolbox" and the "MATLAB Bioinformatics Toolbox" to perform clustering. Using the feature vectors as in Table 1, we generated the phylogenetic tree using the functions "pdist" with default parameters (Euclidean), and "seqlinkage" with "single" 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 [37], the Molecular Evolutionary Genetics Analysis software.

To determine topological correlation between the phylogenetic trees, we measured the congruence by using normalized Robinson-Foulds (nRF) [38] values ranging from 0 to 1. A score of 0 indicates that the trees under investigation are congruent, whereas a score of 1 indicates no congruence, and lower nRF scores indicate a high level of congruence between two trees. The comparison of phylogenetic trees is accomplished using expressions in Newick format. We converted the nRF score into a percentile with the $Accuracy = 1 - nRF*100$ equation. Thus, the percentile value we obtained expressed the frequency of edges in target tree found in

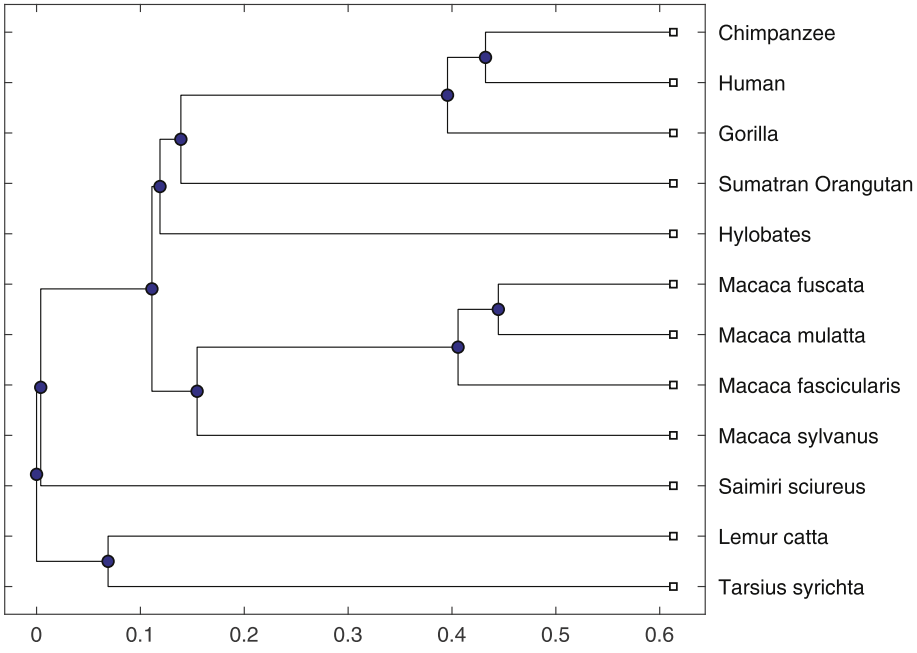

Fig. 2. Phylogenetic tree generated by the proposed method.

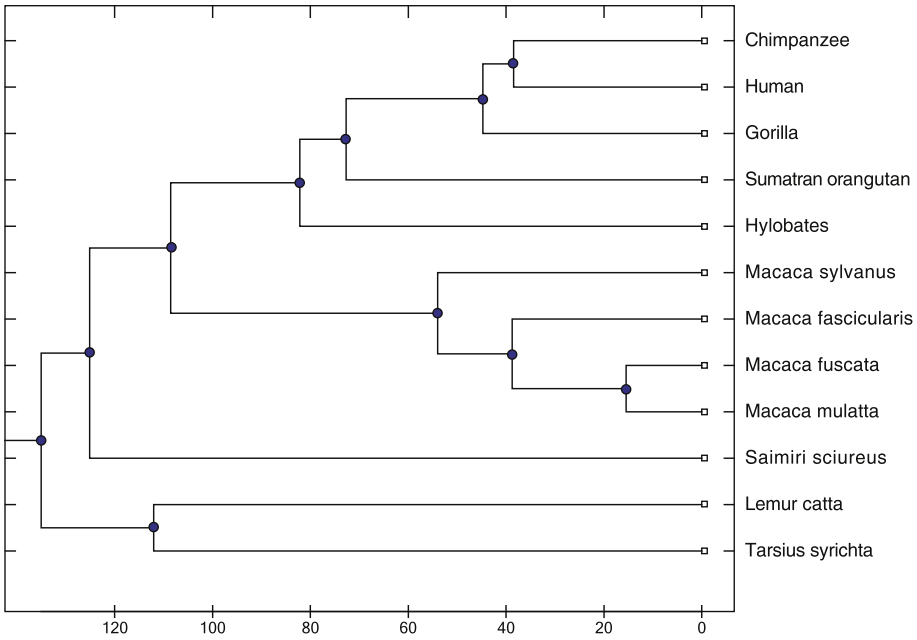

Fig. 3. Phylogenetic tree generated by MEGA7 based on ClustalW alignment and the UPGMA method.

the reference. The success of our method compared to the reference tree from MEGA is expressed as “accuracy rate".

3. Results

3.1. Data description

We applied our method to three DNA datasets and compared the results with the reference trees generated by MEGA software. We chose datasets of different sequence numbers and lengths. We created a data set ourselves and compared it with the reference tree. We also used two additional datasets used by previous researchers to evaluate the results. In addition to there datasets, we used a dataset with known reference phylogenies from benchmark data for phylogeny reconstruction from AFProject [39] which is a free service for objective performance comparison of alignment-free sequence comparison tools on different datasets. The datasets and results are given below according to their length.

Table 4
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 5
Top-*k* *n*-gram accuracy and running time results for 16S ribosomal DNA of 13 bacteria.

n-gram	top-k (percent)	Accuracy	Comput. Time
4	15	91	0.1461
4	17	91	0.1517
14	1	91	0.2470
16	1	91	0.2494
17	1	91	0.2517

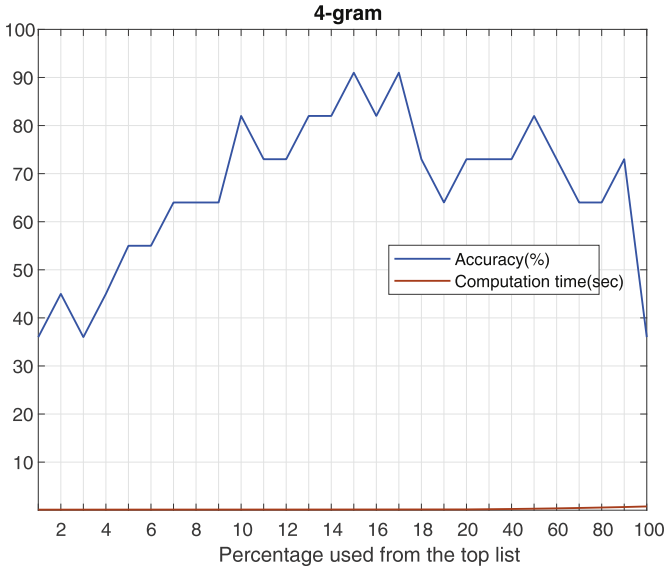

Fig. 4. Accuracy-computation time graph of 4-gram that produces the best results.

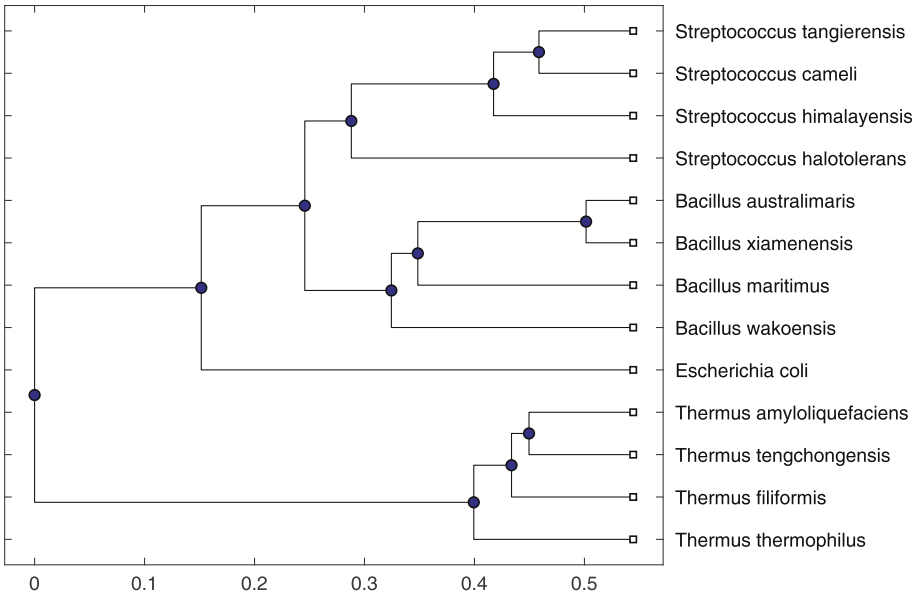

Fig. 5. Phylogenetic tree generated by the proposed method.

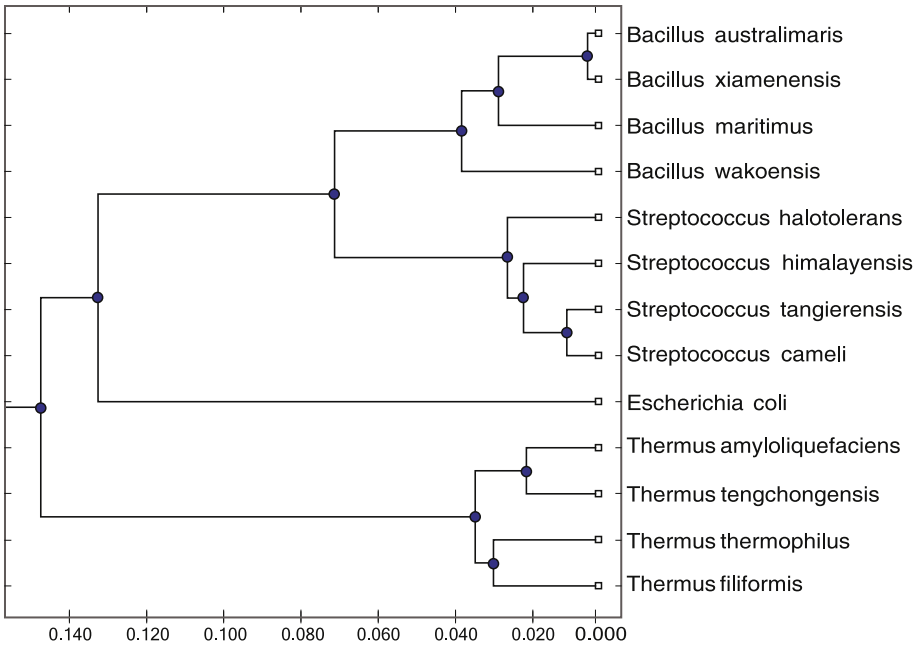

Fig. 6. Phylogenetic tree generated by MEGA7 based on ClustalW alignment and the UPGMA method.

3.2. Implementation

We used different *n*-grams and top-k values in the implementation. We ran our method from 3-gram to 20-gram; at higher values, calculation costs and runtimes become excessive. We used top-k values by increasing one-by-one by predicting that we would find the accuracy value we sought in the first 20 percentage. We used 10-to-10 increments between 20 and 100 to see the higher percentages. The source code of the method was run on a server using Intel Xeon 3.10 GHz CPU, 16 GB RAM and the Windows Server 2012 R2 operating system.

3.2.1. NADH dehydrogenase subunit 4 genes

First, we used the NADH dehydrogenase subunit 4 genes of 12 species of 4 different groups of primates. The dataset consists of 4 species of OldWorld monkeys, one species of New World monkeys, two species of prosimians, and five species of hominoids. All the sequences were obtained from the NCBI database and are listed in Table 2, whose lengths are between 893 and 896 base pairs. They were previously reported and used in their methods by Hayasaka et al. [40], and subsequently used by Zhang [41,42], Qi et al. [11] and Chen et al. [43].

We applied our method from 3-gram to 20-gram, and from top-1 percentage to top-100 percentage. Some results reaching the

Table 6
The 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

Table 7
Top-*k* *n*-gram accuracy and running time results for the whole mitochondrial genome of 18 eutherian mammals.

n-gram	top-k (percent)	Accuracy	Comput. Time
13	4	81	16.2565
12	5	81	21.9551
13	2	75	8.8509
7	8	75	12.8548
7	9	75	15.2619

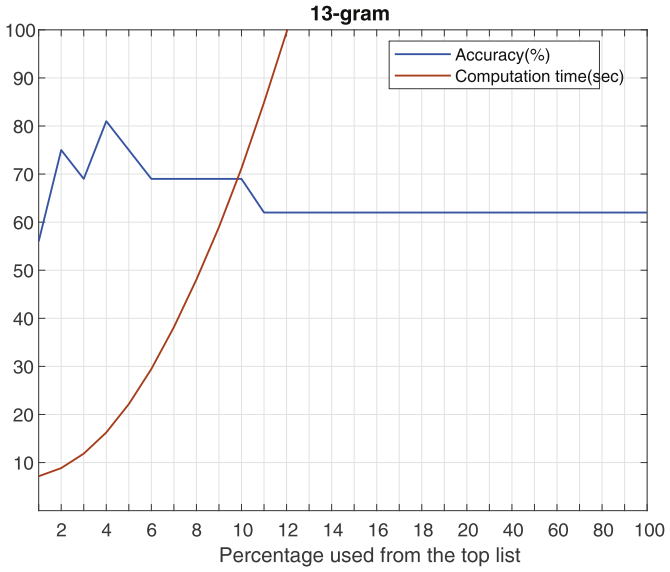

Fig. 7. Accuracy-computation time graph of 13-gram that produces the best results.

highest accuracy value are given in Table 3. There are other situations that reach the highest accuracy in the range in which the tests are performed. But the aim here is to identify the situation that solves the highest accuracy in the shortest calculation time. The accuracy-computation time graph of the *n*-gram producing the best results is given in Fig. 1. The accuracy value was converted to a percentage based on the nRF score. The graphs of all other top-*k* *n*-grams are given in the supplementary material.

After calculating the similarities between the species from the feature vectors, we generated the phylogenetic tree based on this similarity. In Fig. 2, we presented the phylogenetic tree generated by our method, and we also presented the reference phylogenetic tree obtained by MEGA7 based on ClustalW alignment and the UPGMA method in Fig. 3.

3.2.2. 16S ribosomal DNA of bacteria

Second, we chose 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 highly similar sequences, along with well-separated sequences. All sequences were selected from the NCBI database and are listed in Table 4. The sequence lengths are between 1509 and 1541 bases.

We also applied our method for this data set from 3-gram to 20-gram and from top-1 percentage to top-100 percentage. Some results reaching the highest accuracy value are given in Table 5. The accuracy value was converted to percentage based on nRF score. The accuracy-computation time graph of the *n*-gram producing the best results is given in Fig. 4. The graphs of all other top-*k* *n*-grams are given in the supplementary material.

After calculating the similarities between the species from the feature vectors, we generated the phylogenetic tree based on this similarity. In Fig. 5, we presented the phylogenetic tree generated by our method, and we also presented the reference phylogenetic tree obtained by MEGA7 based on ClustalW alignment and the UPGMA method in Fig. 6.

3.2.3. Whole mitochondrial genomes of 18 eutherian mammals

Whole mitochondrial genomes include abundant genetic information of 18 eutherian mammals that has been used frequently in recent years [44]. All of the sequences are obtained from the NCBI database and listed in Table 6, whose lengths are between 16,295 and 17,019 bases.

We applied our method for this dataset from 3-gram to 20-gram and from top-1 percentage to top-100 percentage. Some results reaching the highest accuracy value are given in Table 7. The accuracy value was converted to percentage based on the nRF score. The accuracy-computation time graph of the *n*-gram that producing the best results is given in Fig. 7. The graphs of all the other top-*k* *n*-grams are given in the supplementary material.

After calculating the similarities between the species from the 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 alignment and UPGMA method in Fig. 9.

3.2.4. Complete mtDNA from 25 fish species of the suborder Labroidei

In addition to the above datasets, we used a DNA sequence dataset with known reference phylogenies from benchmark data for phylogeny reconstruction from AFProject [39] which is a free service for objective performance comparison of alignment-free sequence comparison tools on different datasets. AFproject is a collaborative project to benchmark and evaluate software tools and methods for alignment-free sequence analysis in different applications. The web site of the project provides a variety benchmark sequences, and the output of alignment-free methods on these sequences can be uploaded to the server. Thus, the achievement of these methods can be compared to a large number of other alignment-free methods, among them the state-of-the-art methods.

To evaluate our method, we downloaded a dataset of 25 mitochondrial genomes from different fish species of the suborder Labroidei [45] from the AFproject server that are relevant for our

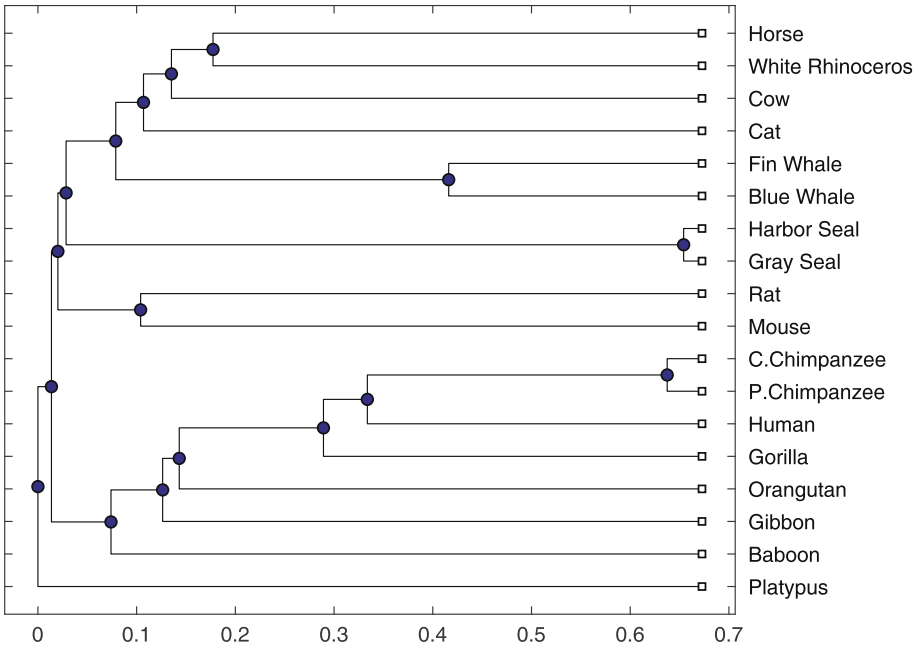

Fig. 8. Phylogenetic tree generated by the proposed method.

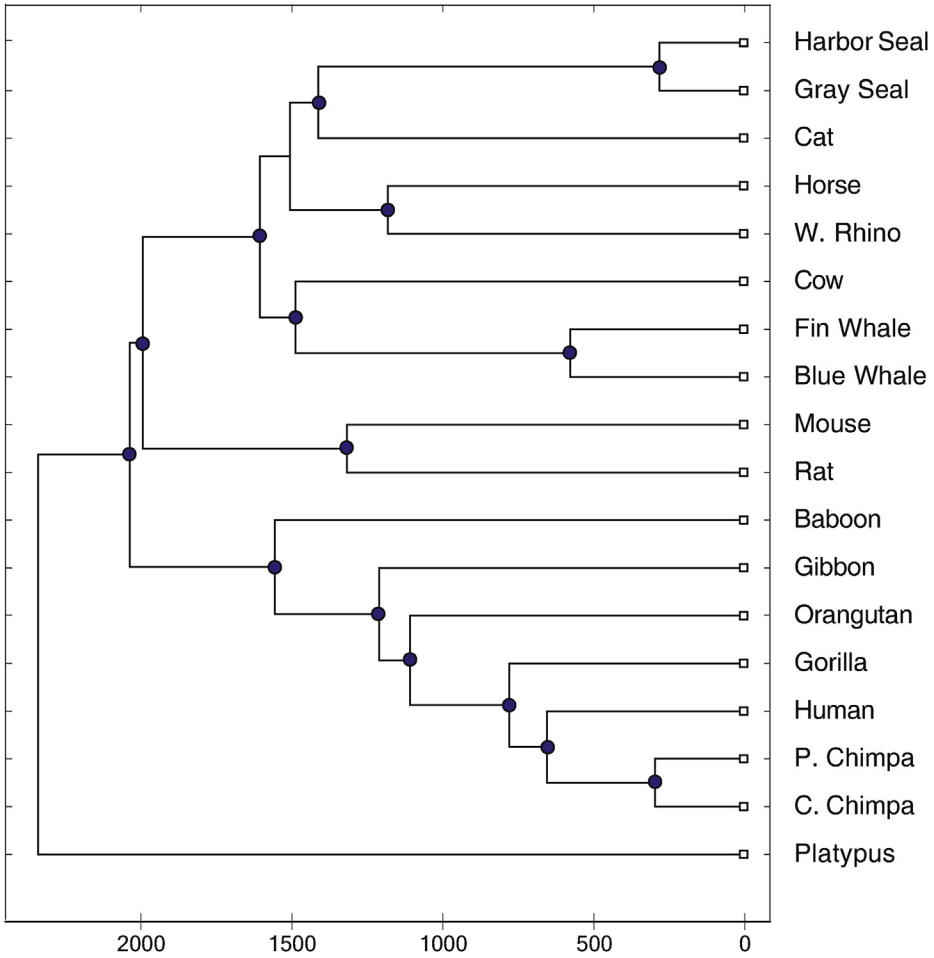

Fig. 9. Phylogenetic tree generated by MEGA7 based on ClustalW alignment and the UPGMA method.

Table 8
Performance comparison between proposed method results and alignment-based methods.

	Computation times (sec)		
Datasets	Top-*k* *n*-gram	ClustalW	Muscle
Table 2	0.19	6.60	2.02
Table 4	0.15	17.70	5.85
Table 6	16.26	4528.05	2877.58

study.This dataset has been previously used by developers of alignment-free methods to test their methods. And it is similar in length and type to the data set we provided in 3.2.3. Therefore, the *k* and *n* values are selected from the values in Table 7.

We ran our method on the above dataset and uploaded the obtained newick tree to the AFproject server for evaluation. The AFproject server compares the obtained tree to their trusted reference tree of the relevant data set under the nRF metric. On the AFproject server, the benchmark results are ranked in order of increasing nRF distance, in a word in order of decreasing quality. As a result, among the 90 methods listed with 18 different accuracy degrees, our method ranks 6th with 68% accuracy rate for this dataset. The benchmark result can be viewed on the AFProject website.

3.3. Performance comparison with alignment-based methods

Different lengths of DNA sequences is a big challenge for similarity analysis methods. Some of these methods cannot effectively deal with the short sequences and long sequences due to their weak characterization capabilities [2]. Word-based methods are more successful in this regard, but complexity is also a serious problem in these methods. In our method, we used frequently repeating high-frequency *n*-grams instead of all *n*-grams in the sequence, which removed the burden of using all *n*-grams and enabled us to use the essence of the DNA sequence in characterization. When Table 8 is examined, the ability of the method that can show success in sequences of different lengths can also be observed. The computation times given for the ClustalW and Muscle alignment-based methods give the MEGA7 program analysis time of the sequences.

4. Conclusion

This paper introduces a different approach to DNA sequence similarity analysis performed on the basis of similarity calculations. In the field of computer science, *n*-gram is a study topic used in similarity calculations of texts for classification, recognition, etc. Top-k *n*-grams of a text or a document consist of the set of *n*-grams with the highest repetition count. We proposed an alignment-free DNA similarity analysis approach based on top-*k* *n*-gram matches, with the prediction that common repeating DNA subsections would indicate high similarity between DNA sequences. With this approach, we used frequently repeating high-frequency *n*-grams instead of all *n*-grams in the sequence, which removed the burden of using all *n*-grams and enabled us to use the essence of the DNA sequence in characterization. In this study we have shown that a certain quantity of frequently recurring common sequence patterns have the power to characterize DNA sequences. In addition, *n*-gram, which is a word-based method, has gained significant advantage in computation time and accuracy, although it includes intense grams comparison. The phylogenetic relationships revealed by our method showed that our trees coincided almost completely with the results of MEGA, which is based on sequence alignment. In addition, it was seen that when we applied the parameters that we suggested to a different dataset of similar length with a data set we used, it showed similar success.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmgm.2020.107693.

References

[1] S.T.F. Wang, Y. Qiu, X. Liu, Bilateral similarity function: a novel and universal method for similarity analysis of biological sequences, J. Theor. Biol. 265 (2) (2010) 194–201.
[2] X. Jin, Q. Jiang, Y. Chen, S. 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.
[3] 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.
[4] 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.
[5] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. U.S.A. 85 (8) (1988) 2444–2448.
[6] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1) (1981) 195–197.
[7] M. Rinku, A. Neeru, A graph theoretic model for prediction of reticulation events and phylogenetic networks for dna sequences, Egyptian Journal of Basic and Applied Sciences 3 (3) (2016) 263–271.
[8] 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.
[9] 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.
[10] N. Jafarzadeh, A. Iranmanesh, C-curve: a novel 3d graphical representation of dna sequence based on codons, Math. Biosci. 241 (2) (2013) 217–224.
[11] 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.
[12] J.F. Yu, J.H. Wang, X. Sun, Analysis of similarities/dissimilarities of dna sequences based on a novel graphical representation, Match-Communications in Mathematical and in Computer Chemistry 63 (2) (2010) 493–512.
[13] 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.
[14] L. Shi, H. Huang, DNA Sequences Analysis Based on Classifications of Nucleotide Bases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, ISBN 978-3-642-27866-2, pp. 379–384.
[15] S. Vinga, J. Almeida, Alignment-free sequence comparison - a review, Bioinformatics 19 (4) (2003) 513–523.
[16] 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.
[17] A. Zielezinski, S. Vinga, J. Almeida, W.M. Karlowski, Alignment-free sequence comparison: benefits, applications, and tools, Genome Biol. 18 (1) (2017) 186.
[18] B.E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, Proc. Natl. Acad. Sci. Unit. States Am. 83 (14) (1986) 5155–5159.
[19] H.U. Osmanbeyoglu, M.K. Ganapathiraju, N-gram analysis of 970 microbial organisms reveals presence of biological language models, BMC Bioinf. 12 (2011).
[20] M.K. Ganapathiraju, A.D. Mitchell, M. Thahir, K. Motwani, S. Ananthasubramanian, Suite of tools for statistical n-gram language modeling for pattern mining in whole genome sequences, J. Bioinf. Comput. Biol. 10 (6) (2012).
[21] M.R. Kantorovitz, G.E. Robinson, S. Sinha, A statistical method for alignment-free comparison of regulatory sequences, Bioinformatics 23 (13) (2007) i249–i255.
[22] Q. Dai, Y. Yang, T. Wang, Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison, Bioinformatics 24 (20) (2008) 2296–2302.
[23] M. Takahashi, K. Kryukov, N. Saitou, Estimation of bacterial species phylogeny through oligonucleotide frequency distances, Genomics 93 (6) (2009) 525–533.

[24] K. Song, J. Ren, G. Reinert, M. Deng, M.S. Waterman, F. Sun, New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing, Briefings Bioinf. 15 (3) (2013) 343–353.
[25] H.H. Huang, C. Yu, Clustering dna sequences using the out-of-place measure with reduced n-grams, J. Theor. Biol. 406 (2016) 61–72.
[26] T.J. Wu, Y.H. Huang, L.A. Li, Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences, Bioinformatics 21 (22) (2005) 4125–4132.
[27] S.R. Jun, G.E. Sims, G.A. Wu, S.H. Kim, Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution, Proc. Natl. Acad. Sci. Unit. States Am. 107 (1) (2010) 133–138.
[28] Q. Li, Z. Xu, B. Hao, Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations, J. Biotechnol. 149 (3) (2010) 115–119. Advanced Methods in Molecular Systems Biology.
[29] S. Vinga, A.M. Carvalho, A.P. Francisco, L.M.S. Russo, J.S. Almeida, Pattern matching through chaos game representation: bridging numerical and discrete data structures for biological sequence analysis, Algorithm Mol. Biol. 7 (1) (2012) 10.
[30] A. Tomović, P. Janičić, V. Kešelj, n-gram-based classification and unsupervised hierarchical clustering of genome sequences, Comput. Methods Progr. Biomed. 81 (2) (2006) 137–153.
[31] B. Chor, D. Horn, N. Goldman, Y. Levy, T. Massingham, Genomic dna k-mer spectra: models and modalities, Genome Biol. 10 (10) (2009) R108, https://doi.org/10.1186/gb-2009-10-10-r108.
[32] M. Alhanahnah, Q. Lin, Q. Yan, N. Zhang, Z. Chen, Efficient signature generation for classifying cross-architecture iot malware, in: 2018 IEEE Conference on Communications and Network Security (CNS), 2018, pp. 1–9.
[33] A. Wagner, V. Bicer, T. Tran, R. Studer, Holistic and compact selectivity estimation for hybrid queries over rdf graphs, in: P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, et al. (Eds.), The Semantic Web – ISWC 2014, Springer International Publishing, Cham, 2014, ISBN 978-3-319-11915-1, pp. 97–113.
[34] B.Y. Cheng, J.G. Carbonell, J. Klein-Seetharaman, Protein classification based on text document classification techniques, Proteins 58 (4) (2005) 955–970, https://doi.org/10.1002/prot.20373.
[35] D. Tauritz, Application of N-Grams, Department of Computer Science, 2002.
[36] D.Z. Wang, L. Wei, Y. Li, F. Reiss, S. Vaithyanathan, Selectivity estimation for extraction operators over text data, in: 2011 IEEE 27th International Conference on Data Engineering, 2011, pp. 685–696.
[37] 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.
[38] D. Robinson, L. Foulds, Comparison of phylogenetic trees, Math. Biosci. 53 (1) (1981) 131–147.
[39] A. Zielezinski, H.Z. Girgis, G. Bernard, C.A. Leimeister, K. Tang, T. Dencker, et al., Benchmarking of alignment-free sequence comparison methods, Genome Biol. 20 (1) (2019) 144.
[40] K. Hayasaka, T. Gojobori, S. Horai, Molecular phylogeny and evolution of primate mitochondrial-dna, Mol. Biol. Evol. 5 (6) (1988) 626–644.
[41] Y.S. Zhang, A simple method to construct the similarity matrices of dna sequences, Match-Communications in Mathematical and in Computer Chemistry 60 (2) (2008) 313–324.
[42] Y.S. Zhang, W. Chen, New invariant of dna sequences, Match-Communications in Mathematical and in Computer Chemistry 58 (1) (2007) 197–208.
[43] 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.
[44] X. Jin, R.C. Nie, D.M. Zhou, S.W. Yao, Y.Y. Chen, J.F. Yu, et al., A novel dna sequence similarity calculation based on simplified pulse-coupled neural network and huffman coding, Phys. Stat. Mech. Appl. 461 (2016) 325–338.
[45] C. Fischer, S. Koblmüller, C. Gülly, C. Schlötterer, C. Sturmbauer, G.G. Thallinger, Complete mitochondrial dna sequences of the threadfin cichlid (petrochromis trewavasae) and the blunthead cichlid (tropheus moorii) and patterns of mitochondrial genome evolution in cichlid fishes, PloS One 8 (6) (2013).