Evolving Systems (2020) 11:637–646
https://doi.org/10.1007/s12530-019-09306-4

ORIGINAL PAPER

Power spectrum and dynamic time warping for DNA sequences classification

Abdesselem Dakhli[1] · Chokri Ben Amar[2]

Received: 13 August 2018 / Accepted: 29 September 2019 / Published online: 22 October 2019

Abstract

Similarity and alignment and are often used to classify DNA sequences. We have developed a new classifier to classify DNA sequence. First, our approach is used to extract the features of DNA strands. Second, the goal of our approach is to classify DNA strands according to the similarity elaborated by the alignment. Frequently, the performance of the classification of DNA sequences depends on the method that allows to extract the characteristics and calculation of the genomic similarity. Particularly, our approach consists of three different methods for improving the classification of the DNA sequences. This paper presents a new approach of classification of DNA sequence based on dynamic time warping (DTW) method. First, the binary indicator is used to code each nucleotide and the power spectrum is used to extract the characteristics. Secondly, the DNA sequence similarity matrix is evaluated by the dynamic temporal Warping. Third, pairwise comparison is used to classify DNA strands. Our approach solves the complex problem of presentation and structure of different groups of organisms. The experimental results of our classifier obtained are compared with other approaches based on the alignment and similarity of the DNA sequences. These results showed that our approach outperformed other approaches in terms of classification and running time. Here is a summary of the main contributions of this article: (1) Convert nucleotides from DNA sequences by applying binary coding. (2) Using power spectrum our approach extracts the characteristics of DNA sequences. (3) Elaborate the similarity matrix of the DNA strand signal by the Dynamic Time Warping method. (4) Use pairwise comparison to classify DNA sequences. The approach developed is efficient to solve the problems of classification of DNA sequences.

Keywords DNA sequences · Power spectrum · Dynamic time warping · Binary · Pairwise comparison · Discrete Fourier transform

1 Introduction

Bioinformatics studies the classification of DNA strands as a fundamental problem in modern genomics. Methods for classifying DNA sequences can be divided into three major types.

The first category is the feature-based classification, which converts a DNA sequence into vector functionality and then uses classical classification methods to classify the DNA sequences. The second type is the classification based on the distance between the DNA strands. Similarity was used to group the DNA sequences. The similarity assessment is done using a distance function that computes the identity between the sequences. The third category uses statistical methods and models to classify DNA strands, for example Hidden Markov Model (HMM) (Xing et al. 2010; Stranneheim et al. 2010; Muller and Koonin 2003; Kawulok and Deorowicz 2015; Ames et al. 2013; Wood and Salzberg 2014; Xu and Jackson 2019). Following our study, we will draw up some representative methods that concern the three categories mentioned.

This paper presents a new approach of classification of DNA sequence based on dynamic time warping (DTW) method.

In this study, we compare our approach with other methods using similarity and alignment to classify DNA strands, such as Carma, Megan, MetaPhyler and Mg-Rast. In the next

✉ Abdesselem Dakhli
abdesselemdakhli@gmail.com

Chokri Ben Amar
chokri.benamar@ieee.org

[1] Hail University, Community College, Hail, Kingdom of Saudi Arabia

[2] REGIM: Research Groups on Intelligent Machines, National Engineering School of Sfax (ENIS), University of Sfax, 3038 Sfax, Tunisia

work, we will compare our approach with classifiers that use short subsequences (k-mers) to classify DNA sequences, such as CoMeta, LMAT and Kraken.

The general objective of our work is to build a classifier to classify DNA sequences. This classifier is based on binary coding, power spectrum and DTW method.

The main contributions of our approach consist at the beginning, the use of binary coding to convert the DNA sequences. Next, our approach extracts the characteristics of the DNA sequences by applying the power spectrum method, elaborates the matrix similarity of the DNA strand signal by the DTW method. Finally, pairwise comparison is used to classify DNA sequences. The approach developed is effective in solving the problems of classifying DNA sequences.

The approach to be developed contains three phases. The first phase is used to encode the DNA strands, which consists of two substeps; binary coding and signal processing of DNA. The second phase makes it possible to calculate the similarity matrix of the genomic signal. The third phase exploits the classification of DNA strands using the pairwise comparison. In this study, the DTW method is applied to solve the similarity of the DNA strands. Our approach uses Power Spectrum (PS) to build Feature Extraction and Time Warping algorithm and pair wise comparison to group DNA sequences.

This study is organized as follows: Sect. 1 provides a review of the literature of major research. Section 2 outlines our method to develop. Section 3 presents the experimental results of our approach. Section 4 concludes with a conclusion and discussion.

2 Background

The classification of DNA sequences has been treated by several works. H. Stranneheim et al. proposed an approach classifies DNA sequences (FACS) in a precise and fast manner. This approach is used to classify the DNA sequences (Stranneheim et al. 2010). It allows classifying the DNA in a precise and a fast way. It was validated using a synthetic metagenome of DNA sequences. The metagenome database was applied to show that the approach achieves a precision comparable to BLAT but is at least 21 times faster in the classification of DNA strands.

The Principal Component Analysis (PCA) was applied by Muller and Koonin (Muller and Koonin 2003) in 2003 to group a DNA database. These are converted into document vectors that represent the content of their word; the PCA then develops groups of DNA strands distributed by Gauss. The approach is used to classify multiple sets of DNA sequence data. This approach is powerful for grouping exons and introns with up to 96% accuracy.

There are other approaches that use short subsequences (k-mers) to classify DNA sequences. For example, the methods CoMeta (Classification of metagenomes) (Kawulok and Deorowicz 2015), LMAT (Livermore Metagenomics Analysis Toolkit) (Ames et al. 2013), Kraken (Wood and Salzberg 2014) and Machine learning (Xu and Jackson 2019; Lo Bosco and Di Gangi 2017) to classify DNA strands in a fast manner unlike BLAST-based approaches, where each reference DNA strand is compared with the query containing the unknown DNA sequence. The results obtained prove that the performance of this approach depends on the k value of the short subsequences (k-mers) so the limit of the performance of this approach depends on the length of the sub-sequences and the DNA database. The Kraken, LMAT and CoMeta programs use the k-mer counting principle. The complexity of these programs depends on k-mer count, DNA sequence size, and the size of the DNA database. In Kraken, such a bias would create many large search ranges, which would require more time to discover.

The Wavelet Neural Networks (WNN) (Dakhli and Bellil 2016) was used to classify DNA Sequence Classification applying the Genetic Algorithms and the Least Trimmed Square. The experimental results are showed that the WNN model outperformed the other models in terms of both the clustering results and the running time.

The artificial intelligence (Xu and Jackson 2019) was applied to classify the Nucleotide sequence data corresponding to the length of 822 bp, collected from the Immuno Polymorphism Database, were compressed to 2-dimensional representation and were plotted. Profiles of the two-dimensional plots indicate that the alleles can be classified as clusters are formed. The two-dimensional plot of HLA-A DNAs gives a clear outlook for characterizing the various alleles.

Similarly, there are other alignment-based approachs to classify the DNA strands. These approaches are Carma, Megan, MetaPhyler and Mg-Rast (Bazinet and Cummings 2012; Liu 2010). They use alignment and similarity to classify the DNA strands.

CARMA is an approach that classifies the taxonomy of unassembled and assembled DNA strands that has been adjusted to work with HMMER3 and BLAST homology searches. This approach shows low taxonomic predictions (at the same sensitivity) as other BLAST methods (Bazinet and Cummings 2012; Liu 2010). The classification efficiency of CARMA was verified on short DNA with known taxonomic origins. This method permits a low classification rate when the DNA strands to be classified have a long size.

The MEGAN program (Bazinet and Cummings 2012; Liu 2010) uses the lowest common ancestor (LCA) method. This program is based on the Basic Local Alignment Search Tool (BLAST). All BLAST hits with a score close to the best hit are ranked. The DNA strands are then grouped together by calculating the LCA of all database sequences. This method selects

DNA strands with ambiguous occurrences at higher taxonomic levels to improve classification performance (Miyake et al. 2018). This method gives a low classification rate when the DNA strands to be grouped have a significant length.

The MetaPhyler program is applied to group the small strands of DNA. The marker genes are applied as a taxonomic reference. The results demonstrate that MetaPhyler outperforms other methods commonly used in this context (CARMA, Megan and PhymmBL) (Bazinet and Cummings 2012; Liu 2010).

The MG-RAST program groups the input DNA strands containing genes encoding particular proteins(GEPs), and each gives a specific taxon. All DNA fragments of queries from a random sample do not contain GEPs), so this program typically has fewer overall strands than other methods.

3 Methods

The approach to be developed in this study classifies DNA sequences. This approach is based on the DTW algorithm and the Power Spectrum (PS-DTW). The PS is applied to construct the features extraction of the DNA sequence and DTW is used to group the features extraction of the DNA.

3.1 The alphabet representation of a DNA

The strand of DNA is composed by the following nucleotides: A, C, G and T. Each strand can be biologically presented by an alphabetic representation constituting a chain of letters A, C, G and T. Assume X = {×0, x1, x2,…, xm}. Where xi ∈ {A, C, T, G} and m indicates the size of a DNA sequence.

3.2 The genomic signal of a DNA sequence

Our approach attempts to classify the DNA strands of the species. The nucleotides in each DNA sequence are respectively: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T), where each organism is defined by its DNA sequence (bai Arniker et al. 2012; Mavromatis et al. 2007).

Binary coding transforms the DNA strand into a genomic signal. The latter is used to obtain the signal spectrum of the DNA strand by setting 0 or 1 for no or the existence of a specific nucleotide (A, C, T, G) at the DNA strands (bai Arniker et al. 2012; Lei et al. 2012). For a DNA sequence x [7] = [A A A C G T T], the following coding is obtained: x [7] = [1000 10001000 0100 0010 0001 0001]. The codification obtained from a DNA strand can be processed using mathematical models, for example the discrete Fourier transform (DFT) is used to obtain a complex number f (x):

$$f(x) = \sum_{n=0}^{N-1} X_e(n)\, e^{-j\pi n/N}, \quad k = 0, 1, 2, \ldots, N-1. \qquad (1)$$

We apply a Power Spectrum (PS) approach to analyze and treat the periodicity of DNA strand components. This approach uses periodic nucleotide distributions of DNA sequences to evaluate power spectra at particular and well-defined periodicities. The amplitude of a PS gives the power of a signal on periodic positions. The PS minimizes the background noise that appears very high in the Fourier power spectrum and eliminates spectral leakage, unlike for DFT. Thus, PS captures the invisible periodicities in the sequences. Applying a sliding window, PS develops a precise location of the periodic parts in the DNA strands. We use the PS to discover the hidden intervals in the elements of DNA strands.

This function is easier to apply with the power spectrum sequence than with the original discrete TF. The SP for the frequencies k = 0, 1, 2,…, N−1 is given as follows,

$$SP[k] = |f(x)|^2 \qquad (2)$$

SP [k] was represented by Fig. 1.

3.3 The dynamic time warping (DTW)

The deformation of dynamic time is primarily designed to analyze and process speech (Legrand et al. 2008). The same word has the same importance (the signal has the same form) in the speech of different people, but its timing and deviation are precise to each person. The DTW approach can adjust the offset and timing of signals (Skutkova et al. 2013). This property can be used for DNA processing if the DNA strand is in the form of a genomic signal. The DTW is a general version of the dynamic programming method that determines exactly the pairwise biological sequence alignment problem. The nucleotide position, amplitude and cumulative phase of signals are obtained by the transformation of the time variable. The word of a spoken language is

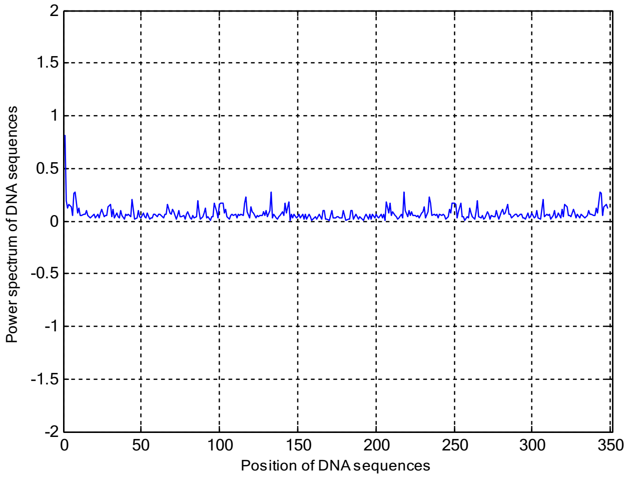


Fig. 1 Signal from a DNA strand obtained by PS

considered a similar pattern of sequences, most often similarly patterned strands of DNA. The nucleotide positions of DNA strands are dynamic; consequently, it is required for using the dynamic programming systems, such as, DTW. This system aligns the sample values according to the minimization of the similarity between the pairs of strand samples. Stretching of one or both signals is performed by repeating the selected samples. The criterion of comparison and iteration of the samples is achieved by the accumulated distances.

In this article (Skutkova et al. 2013), we use the DTW with warung paths pruned. We have adjusted and used the traditional DTW method to discover and remove cells in the DTW matrix that are guaranteed not to lead to alignments that will lead to the optimal path. The DTW algorithm match path theory with global constraints is theoretically shown for our pruning strategy, which is useful for minimizing and reducing the time complexity and computational cost.

The cumulative values of the distance are made using the distances in pairs for each pair of samples according to (3).

$$\begin{aligned} Dis(i,k) &= \min\{Dis(i-1,k-1), \\ &Dis(i-1,k), Dis(i,k-1)\} + dis(i,k) \end{aligned} \tag{3}$$

Where dis is a distance value in pairs and Dis is the cumulative distance. The value of the cumulative distance Dis (i, k) is achieved by the distance dis (i, k) and minimum from the previous values of the cumulative deviations. This set of accumulated distances for each pair of samples forms a table. The buckling of the result series is recovered as a function of the decrease of the return path from the upper right terminal to the lower left terminal.

The final step required for the comparison of DNA strand signals is the comparison of the size and position of the strand signals by DTW. The use of DTW on all DNA strand signals has been performed in the realization of the dendrograms using a cluster analysis.

The resemblance of two matched DNA strand signals is assessed by their normalized Euclidean distance to the length of the aligned signals. The resemblance matrix for the cluster analysis is based on the similarity values evaluated for each pair of signals corresponding to the DTW.

3.4 Proposed algorithmic approach

In this part, we develop a procedure for grouping DNA strands into classes. The details of our procedure are as follows:

Phase 1: Partition the DNA database into two sets of data (Learning and Training DNA Set).
Phase 2: Use binary coding to encode symbolic DNA strands into a digital signal.
Phase 3: Use the discrete Fourier Transform to compute the DNA strand signal.
Phase 4: Convert the digital signal using the Power Spectrum Signal Processing.
Phase 5: Apply the DTW to develop the similarity matrix.
Phase 6: Classify the DNA sequence using the similarity matrix and the pair-wise comparison (a distance is computed between every pair of DNA strands).
Phase 7: Construct of the dendrogram tree.

The following function uses the DTW method to generate the DNA sequences similarity matrix.

```
function  [MatrixDTW]=ComputMatrixDTW(Seq1, Seq1)
% DNA sequence signal Seq1
% DNA sequence signal Seq2
n= length(Seq1);
m= length (Seq2);
MatrixDTW (0,0)=0;
for I=1:n
MatrixDTW (I,1)= MatrixDTW (I-1,1)+c(I,1);
end
for J=1:m
MatrixDTW (1,J)= MatrixDTW (1,J-1)+c(1,J);
end
for I=1:n
for K=1:m
MatrixDTW (I,K)= c(I,K)+min(MatrixDTW (I-1,K), MatrixDTW (I,K-1), MatrixDTW (I-1,K-1))
end
end
end
```

To find the best comparison between two DNA strands, we must discover the path through the grid P = p1,…, pw,…, pj.

pw = (iw,kw) which minimizes the total distance between them. P is called a warping function.

The classification technique to be developed in our work is based on the alignment and the similarity of the DNA strands. During the learning phase, our approach converts the DNA strands applying binary coding and Power Spectrum. In this phase, our approach is applied to distinguish between different groups (classes) using examples of the DNA strands in the training base. Indeed, our system designs a template for each learning strand so each sequence will be known by a signal. Then, our procedure elaborates a similarity matrix applying the DTW algorithm which allows seeking an optimal pairing between each pair of temporal signals of DNA strands. So, this algorithm first develops a nonlinear comparison in the search among all the possible alignments, that which reduces the cumulative cost function. Our approach groups the DNA strands in the following way: at each step, we gather in a class the two strands that have a minimal distance (Figs. 2, 3).

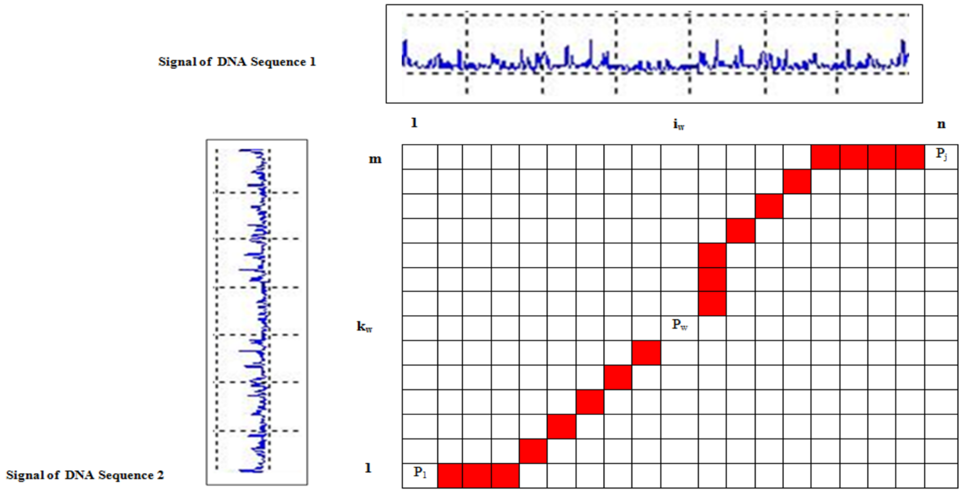


Fig. 2 The similarity matrix using dynamic time warping (DTW)

The classification technique to be developed in our work is based on the alignment and the similarity of DNA strands. During the test phase, a new strand is classified using optimal matching between this sequence and the grouped sequences. Assignment this strand in a group is based on the similarity function. So, the sequence is assigned in a class where the distance is minimal between this sequence and the other sequences within the same class.

To justify the effectiveness of our developed classifier, several assessment measures must be applied, including the accuracy, sensitivity, precision and clustering rates of the DNA strands.

These evaluation criteria are applied to compare our classifier with other competitive classifiers that use the comparison principle to group DNA strands together, for example, Carma, Megan, MetaPhyler and Mg-Rast.

The Precision can be defined as the percentage of correctly ranked instances. The Accuracy (4) is the most intuitive measure of performance and it is clearly a relation between the exactly predicted observation and the set of observations.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

Where FN, TP, FP, and TN indicate the number of false negatives, true positives, false positives, and true negatives, respectively.

The sensitivity (5) is the capability of a test to discover a true positive.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{6}$$

The precision (6) indicates the ratio of the positive predicted positive observations to the total predicted positive observations.

4 Results

This paper uses three datasets PhyloPythia 961 bp, MetaPHYLER 300 bp and FACS 269 bp. Each dataset was randomly selected from Harvard Dataverse (https://dataverse.harvard.edu/dataverse/harvard). The DNA sequence sizes in the three databases are respectively 269 bp, 300 bp and 961 bp. Similarly our approach was simulated on two other databases (16S-AMP-trimmed.fa.zip and 16S-SG-reads.fa.zip) to further validate our classifier (Fiannaca et al. 2018; https://github.com/IcarPA-TBlab/MetagenomicDC/tree/master/data).

The description of the databases used found in Table 1.

Numerical values of accuracy prove the efficiency and performance of our approach. It specifies the correct classification of DNA strands. We compare the PS-DTW (our approach) and the four alignment models, i.e., Carma, Megan, MetaPhyler and Mg-Rast. Our approach for all the DNA strands datasets (FACS269 bp, MetaPHYLER 300 bp and PhyloPythia 961 bp) has high accuracy, which is, respectively, 51.5%, 97.3% and 71.8%. For strands of size 300 bp and 961 bp the accuracy of our approach is higher than that obtained by the other models at all the DNA strands datasets. It is clear that the performance of our approach.

The results obtained (Fig. 4) show the degree of relationship between the organisms and thus make it possible to understand their evolutionary history. Class 1 groups strands (1, 4, 5, 12,…). This grouping shows well the biological relation which exists between the organisms which are identified by their DNA strands. From these results it is found that there are common biological characteristics between the species.

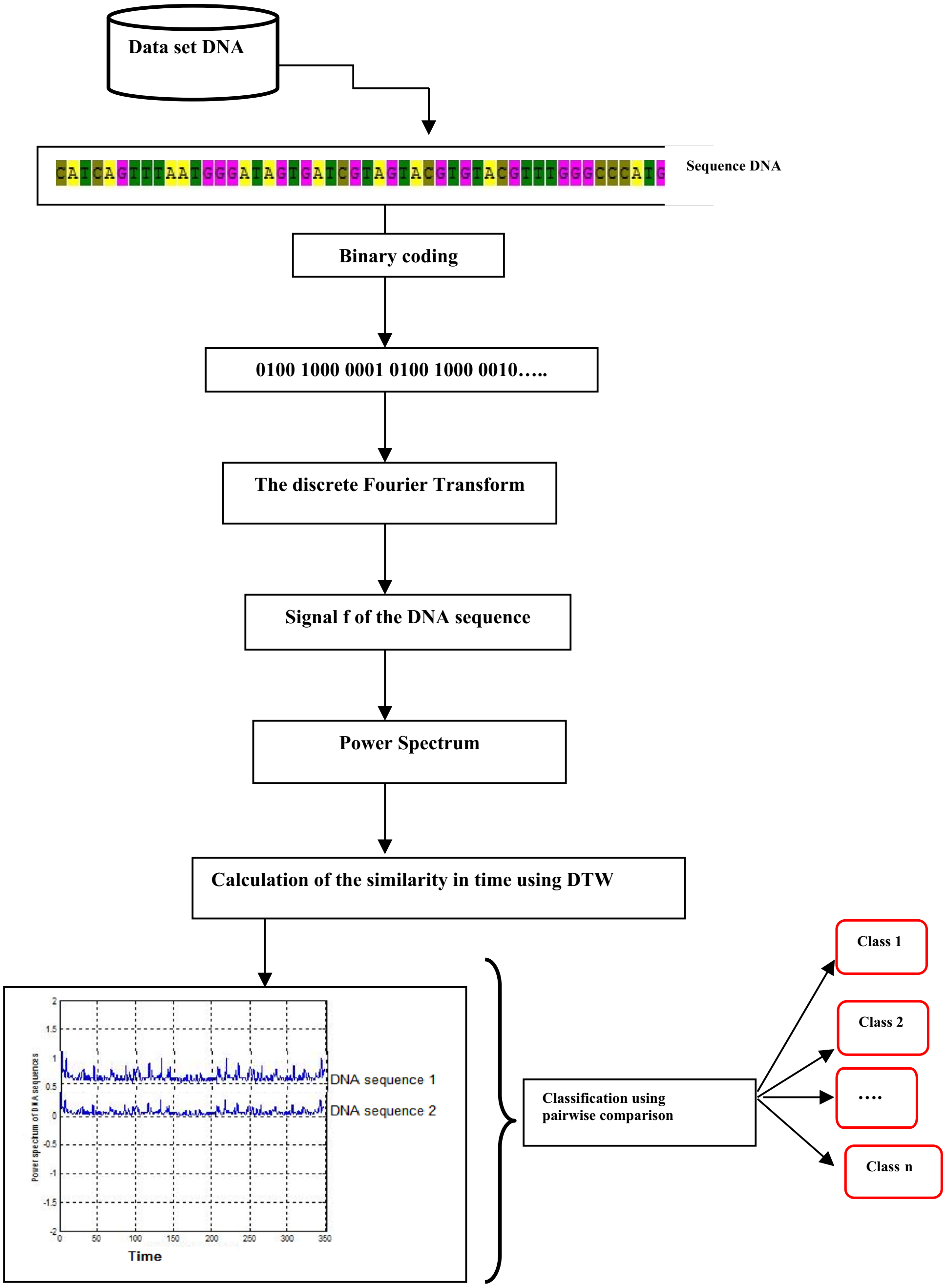

Data set DNA
CATCAGTTTAATGGGATAGTGATCGTAGTACGTGTACGTTTGGGCCCATG
Sequence DNA
Binary coding
0100 1000 0001 0100 1000 0010…..
The discrete Fourier Transform
Signal f of the DNA sequence
Power Spectrum
Calculation of the similarity in time using DTW
Power spectrum of DNA sequences
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
50
100
150
200
250
300
350
Time
DNA sequence 1
DNA sequence 2
Classification using pairwise comparison
Class 1
Class 2
….
Class n

Fig. 3 Proposed approach

The grouped DNA strands have a biological resemblance according to Fig. 5. This similarity can affirm a similar biological role within the cells of living beings. The groupings developed are very positive and prove that the proposed approach can be successfully applied to solve the problem of grouping DNA strands.

The DNA resemblance can biologically improve the current family of DNA strands that are unable to directly affect the considerable volume of these strands. We use another dataset of the DNA sequence of the FACS database to improve and verify the performance of our approach for large sequential datasets.

The instability of a test is its ability to correctly solve cases of DNA sequences. To evaluate it, we will develop the percentage of true positives in the case of DNA sequences In terms of unified accuracy and sensitivity, our approach surpassed all other classifiers, producing a 95.3% sensitivity for the MetaPHYLER dataset (Table 4).

The results obtained prove that our approach (PS-DTW) is efficient for classifying a DNA database. It can access a very good precision. Our classifier is experienced on empirical databases, exposing that accuracy exceeds other techniques in terms of percentage of correct classification of organisms. The results obtained (Tables 2, 3, 4 and 5) show

Table 1 Description of each dataset

	DNA databases				
	FACS (https://dataverse.harvard.edu/dataverse/harvard)	MetaPHYLER (https://dataverse.harvard.edu/dataverse/harvard)	PhyloPythia (https://dataverse.harvard.edu/dataverse/harvard)	SG (Fiannaca et al. 2018; https://github.com/IcarPA-TBlab/MetagenomicDC/tree/master/data)	AMP (Fiannaca et al. 2018; https://github.com/IcarPA-TBlab/MetagenomicDC/tree/master/data)
Sizes	269 bp	300 bp	961 bp	400 bp	300 bp
Number of sequences	105	73	25	100	90
Training	60	40	15	80	70
Test	45	33	10	20	20

the good distribution of DNA in the classes obtained. During the test phase our classifier produces good classification rates.

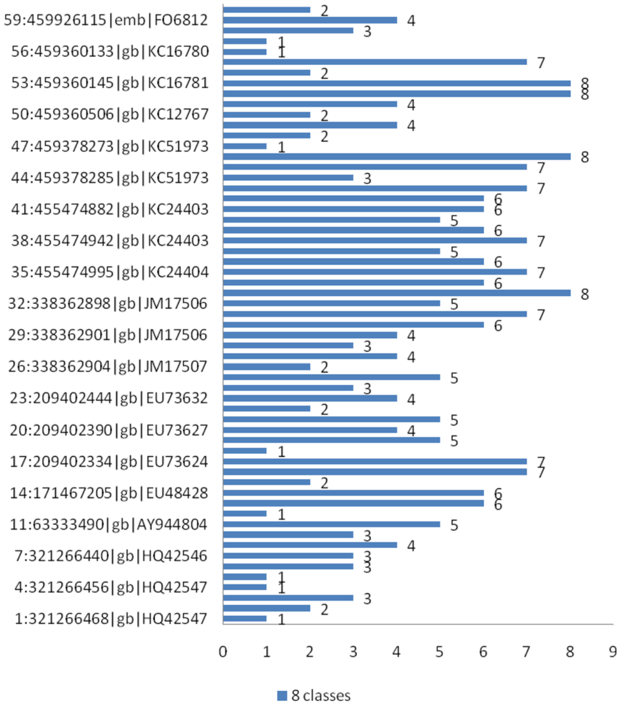
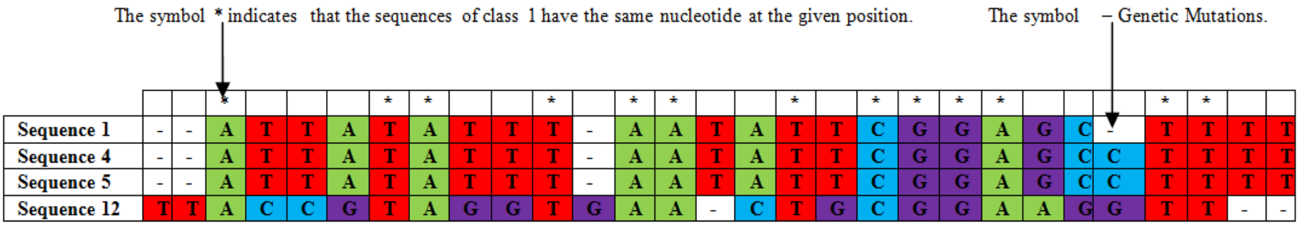


Fig. 4 Graphical representation of DNA groups of FACS database containing 60 sequences

Our approach shortens the running time for all the DNA strands datasets also improves the classification accuracy of the strands of DNA. Table 6 and Fig. 6 show the CPU runtime of all the methods on the four databases of the DNA strands (FACS269 bp, MetaPHYLER 300 bp and PhyloPythia 961 bp). The CPU run time PS-DTW was much shorter than the other models. The CPU run time can affect the time complexity of the Computation of Correspondence between the DNA strands. Thus, the complexity of the models (Carma, Megan, MetaPhy-ler and Mg-Rast) are higher than the our approach. The execution time of the processor grows comparatively with the size of the DNA sequence. The results show that the length of DNA strand prolongs the runtime of the models. The size of the DNA strands has an influence on the execution time; the classification of DNA strands of size equal to 961 bp requires a run time of 21 min (Figs. 7, 8).

Numerical values of accuracy prove the efficiency and performance of our approach. It specifies the correct classification of DNA sequences. We compare the PS-DTW (our approach) and the three algorithms, i.e., CNN (Convolutional Neural Networks), DBN (Deep belief network) and RDP (Ribosomal Database Project) (Fiannaca et al. 2018; https://github.com/IcarPA-TBlab/MetagenomicDC/tree/master/data). Our approach for all the DNA datasets (SG and AMP) has high accuracy, which is respectively 91% and 95%.


The symbol * indicates that the sequences of class 1 have the same nucleotide at the given position.
The symbol – Genetic Mutations.

			*				*	*			*		*	*			*		*	*	*	*				*	*		
Sequence 1	-	-	A	T	T	A	T	A	T	T	T	-	A	A	T	A	T	T	C	G	G	A	G	C	-	T	T	T	T
Sequence 4	-	-	A	T	T	A	T	A	T	T	T	-	A	A	T	A	T	T	C	G	G	A	G	C	C	T	T	T	T
Sequence 5	-	-	A	T	T	A	T	A	T	T	T	-	A	A	T	A	T	T	C	G	G	A	G	C	C	T	T	T	T
Sequence 12	T	T	A	C	C	G	T	A	G	G	T	G	A	A	-	C	T	G	C	G	G	A	A	G	G	T	T	-	-

Fig. 5 Alignement of DNA groups of class 1

Table 2 Percentage of sequence classified of each method on all datasets

		DNA databases		
		FACS 269 bp	MetaPHYLER 300 bp	PhyloPythia 961 bp
	Approaches	Classification accuracy (%)		
	(PS-DTW) our approach	51.5	97.3	71.8
Bazinet and Cummings (2012)	Carma	29.0	93.6	61.3
	Megan	48.4	88.2	62.2
	MetaPhyler	0.2	80.9	0.6
	Mg-Rast	27.1	29.8	70.5

Table 3 The classification results of our approach for all databases of DNA sequences

DNA databases	Size (bp)	#Class	Classification accuracy (%)
FACS	269	60	**51.5**
MetaPHYLER	300	124	**97.3**
PhyloPythia	961	210	**71.8**

Bold values in the "classification" column to present the results obtained by the proposed approach

5 Conclusion

In this paper, a classifier has been developed to group the DNA strands. To obtain a signal for each DNA sequence, the binary coding, the Power Spectrum and the Fourier Transform were used. In our approach, we achieve very good prediction accuracy. To validate our classifier, we applied empirical data sets. The results obtained show that the accuracy of our approach surpasses the other models in

Table 4 Sensitivity of each method on all datasets

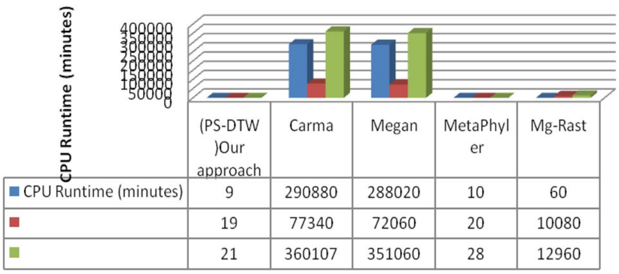
		DNA databases		
		FACS269 bp	MetaPHYLER 300 bp	PhyloPythia 961 bp
	Approaches	Sensitivity (%)		
	(PS-DTW) our approach	50.6	95.3	71.5
Bazinet and Cummings (2012)	Carma	26.7	93.4	59.8
	Megan	42.5	87.9	61.0
	MetaPhyler	0.1	80.7	0.5
	Mg-Rast	25.0	29.7	67.2

Table 5 Precision of each method on all datasets

		DNA databases		
		FACS269 bp	MetaPHYLER 300 bp	PhyloPythia 961 bp
	Approaches	Precision (%)		
	(PS-DTW) our approach	95.5	99.9	98.9
Bazinet and Cummings (2012)	Carma	92.0	99.7	97.4
	Megan	78.1	99.7	98.1
	MetaPhyler	84.0	99.7	83.8
	Mg-Rast	92.4	99.8	95.3

Table 6 Runtime of each method on all datasets

		DNA databases		
		FACS269 bp	MetaPHY-LER 300 bp	PhyloPythia 961 bp
	Approaches	CPU Runtime (min)		
	(PS-DTW) OUR approach	9	19	21
Bazinet and Cummings (2012)	Carma	290880	77,340	360,107
	Megan	288,020	72,060	351,060
	MetaPhyler	10	20	28
	Mg-Rast	60	10,080	12,960



	(PS-DTW)Our approach	Carma	Megan	MetaPhyler	Mg-Rast
🟦 CPU Runtime (minutes)	9	290880	288020	10	60
🟥	19	77340	72060	20	10080
🟩	21	360107	351060	28	12960

Fig. 6 The execution time of our approach with the other methods

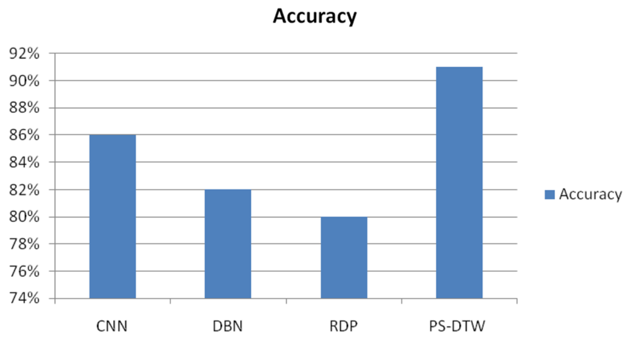


Fig. 7 Accuracy validation of CNN, DBN, RDP and PS-DTW(our approach) classifiers for SG datasets

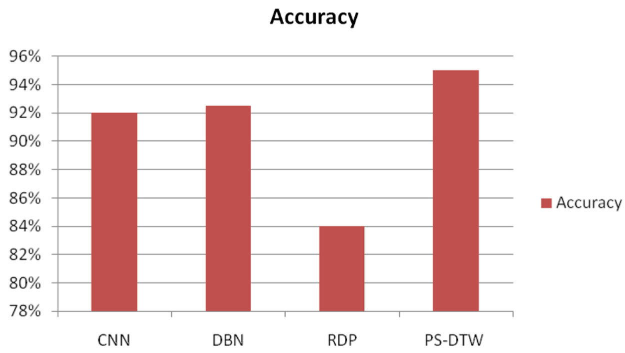


Fig. 8 Accuracy validation of CNN, DBN, RDP and PS-DTW (our approach) classifiers for AMP datasets

terms of the percentage of correct recognition of the DNA sequence. Our developed classifier outperformed the other models (Carma, Megan, MetaPhyler and Mg-Rast) in terms of execution time and classification results. This approach uses biological information to group species into different groups. The classification obtained gives us knowledge about the evolution of species during history. These results were obtained due to two competences presented as follows:

- The efficiency of the binary conversion and the PS approach to reach the signal of DNA strands.
- The competency of the DTW algorithm and the pairwise comparison technique. Using these two means our approach builds a suit number of data groups applying the DNA strands characteristics.

Acknowledgements I want to thank the Research Group on Intelligent Machines (REGIM).

References

Ames S, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE (2013) Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics 29(18):2253–2260. https://doi.org/10.1093/bioinformatics/btt389 **(PMID:23828782.+5)**

bai Arniker S et al (2012) Advanced numerical representation of DNA sequences. In: International conference on bioscience, biochemistry and bioinformatices IPCBEE, vol 31, pp 1–5

Bazinet AL, Cummings MP (2012a) A comparative evaluation of sequence classification programs. BMC Bioinform 13(1):92

Bazinet AL, Cummings MP (2012b) A comparative evaluation of sequence classification programs. BMC Bioinform 13(1):92

Dakhli A, Bellil W, Ben Amar C (2016) Wavelet neural networks for DNA sequence classification using the genetic algorithms and the least trimmed square. Procedia Comput Sci 96:418–427

Fiannaca A, La Paglia L, La Rosa M, Renda G, Rizzo R, Gaglio S, Urso A (2018) Deep learning models for bacteria taxonomic classification of metagenomic data. BMC Bioinforms 19(7):198

Kawulok J, Deorowicz S (2015) CoMeta: classification of metagenomes using k-mers. PloS one 10(4):e0121453

Legrand B et al (2008) Chromosome classification using dynamic time warping. Pattern Recogn Lett 29:215–222

Lei Ch, Yu-Mei Ch, Ding-Cheng C, Xiao-Wen S (2012) DNA Barcodes and species and subspecies classification within genus Carassius. Zool Res 33:463–472

Liu B et al (2010) MetaPhyler: Taxonomic profiling for metagenomic sequences. 2010 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, USA, pp 95–100

Lo Bosco G, Di Gangi MA (2017) Deep learning architectures for DNA sequence classification. In: Petrosino A, Loia V, Pedrycz W (eds) Fuzzy logic and soft computing applications. WILF 2016. Lecture Notes in Computer Science, vol 10147. Springer, Cham

Mavromatis K et al (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods 4(6):495–500

Miyake J et al (2018) Graphical classification of DNA sequences of HLA alleles by deep learning. Hum Cell 31.2:102–105

Muller H-M, Koonin SE (2003) Vector space classification of DNA sequences. J Theor Biol 223:161–169

Skutkova H et al (2013) Classification of genomic signals using dynamic. BMC Bioinform 14(Suppl 10):S1

Stranneheim H et al (2010) Classification of DNA sequences using Bloom filters. Bioinformatics 26(13):1595–1600

Wood DE, Salzberg SL (2014) Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15(3):R46. https://doi.org/10.1186/gb-2014-15-3-r46 **(PMID: 24580807+6)**

Xing Z, Pei J, Keogh E (2010) A Brief Survey on Sequence Classification. ACM SIGKDD Explor 12(1):40–48

Xu C, Jackson SA (2019) Genome Biol 20:76

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