UNIVERSIDAD NACIONAL DE SAN AGUSTÍN

Mineria de Datos

An analysis of alignment-free methods using image textures from DNA sequences

MSc. Vicente Machaca Arceda

August 5, 2020

Overview

Introduction

- Problem
- Objective

Alignment-free methods base on image textures

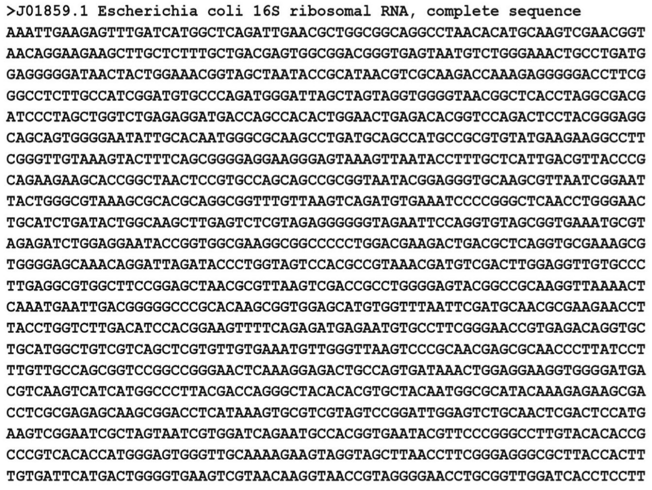
- First-order statistics (FOS)
- Gray Level Co-ocurrence Matrix (GLCM)
- Multi-resolution Local Binary Patterns (MLBP)

Results

- Datasets
- Comparison of FOS, GLCM, LBP and MLBP
- Comparison of mapping functions

Conclusions

Introduction
DNA sequence



```
>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
```

Figure: 16S ribosomal DNA of Escherichia coli with FASTA Format.

Introduction

DNA sequence

The human genome is made of ~**3.2 billions bp** of DNA.
~6.4 billions of nucleotides [1].

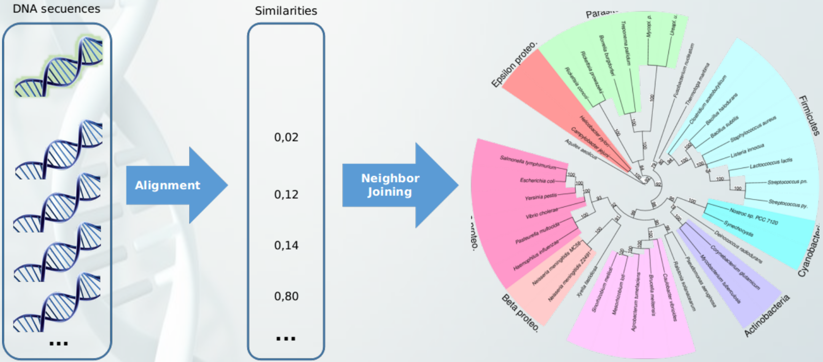
The HIV-1 genome is made of ~**20k bp** of DNA.
Meanwhile, the COVID-19 is made of ~**32k bp** [2].

Overview

- Introduction
  - Problem
  - Objective
- Alignment-free methods base on image textures
  - First-order statistics (FOS)
  - Gray Level Co-ocurrence Matrix (GLCM)
  - Multi-resolution Local Binary Patterns (MLBP)
- Results
  - Datasets
  - Comparison of FOS, GLCM, LBP and MLBP
  - Comparison of mapping functions
- Conclusions

Problem

Phylogenetics steps



DNA secuences

...

Alignment

Similarities

0,02

0,12

0,14

0,80

...

Neighbor Joining

Figure: Steps to visualize phylonetics trees.

Problem

Phylogenetics steps

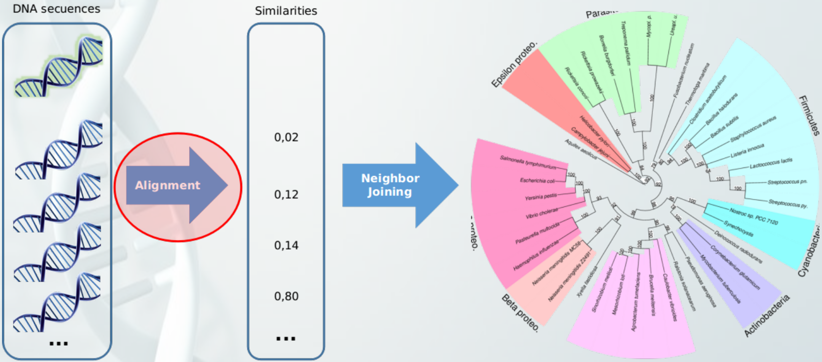


Figure: Steps to visualize phylonetics trees.

Problem

Alignment-based methods

- The most used **alignment-based** method are BLAST and CLUSTALW.

Problem

Alignment-based methods

- The most used **alignment-based** method are BLAST and CLUSTALW.
- They are slow. For example, it take one hour to align 18 sequences of 18k bp.

Problem
Alignment-based methods

- The most used **alignment-based** method are BLAST and CLUSTALW.
- They are slow. For example, it take one hour to align 18 sequences of 18k bp.
- DNA sequences increases every day so **alignment-based** methods get slower every second.

Problem

Alignment-based methods

- The most used **alignment-based** method are BLAST and CLUSTALW.
- They are slow. For example, it take one hour to align 18 sequences of 18k bp.
- DNA sequences increases every day so **alignment-based** methods get slower every second.

Overview

Introduction

- Problem
- Objective

Alignment-free methods base on image textures

- First-order statistics (FOS)
- Gray Level Co-ocurrence Matrix (GLCM)
- Multi-resolution Local Binary Patterns (MLBP)

Results

- Datasets
- Comparison of FOS, GLCM, LBP and MLBP
- Comparison of mapping functions

Conclusions

Objective

Compare **alignment-free** algorithms based on texture descriptors, against CLUSTALW.

- First-Order Statistics **(FOS)** [3].
- Gray Level Co-ocurrence Matrix **(GLCM)** [6].
- Multi-resolution Local Binary Patterns **(MLBP)** [7].

Compare the phylogenetic tree's distances with Robinson Fould [8], and trees' structure with Phylo.io [9].

Overview

Introduction

- Problem
- Objective

Alignment-free methods base on image textures

- First-order statistics (FOS)
- Gray Level Co-ocurrence Matrix (GLCM)
- Multi-resolution Local Binary Patterns (MLBP)

Results

- Datasets
- Comparison of FOS, GLCM, LBP and MLBP
- Comparison of mapping functions

Conclusions

FOS

Mapping function

Each pair of bases have a value from 0 to 15.

$$\alpha = \left\{ \begin{array}{l} AA, AG, AC, AT, GA, GG, GC, GT, \\ CG, CC, CT, CA, TA, TG, TC, TT \end{array} \right\} \quad (1)$$

FOS

Image from DNA

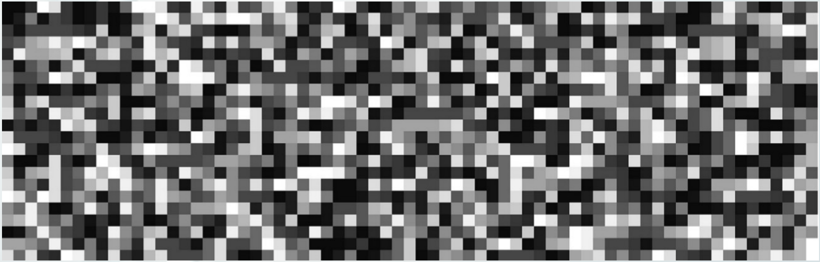


Figure: Textures converted from the DNA sequences of Bacillus maritimus 16S ribosomal DNA.

FOS

Histogram

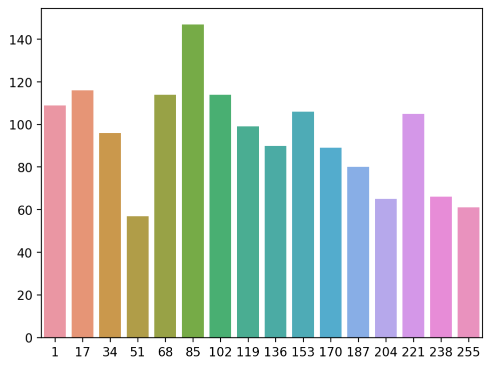


Figure: Histogram of Bacillus maritimus 16S ribosomal DNA.

FOS
Features

From the histogram, the following features are compute:

- $\textbf{\textit{Skewness}} = \sigma^{-3}\sum_{i=0}^{G-1}(i-\mu)^3 p(i)$
- $\textbf{\textit{Kurtosis}} = \sigma^{-4}\sum_{i=0}^{G-1}(i-\mu)^4 p(i) - 3$
- $\textbf{\textit{Energy}} = \sum_{i=0}^{G-1} p(i)^2$
- $\textbf{\textit{Entropy}} = -\sum_{i=0}^{G-1} p(i)lg(p(i))$

Where:

- $p(i) = h(i)/NM$
- $h(i) = histogram$
- *N* and *M are image's width and height.*
- $\mu = \sum_{i=0}^{G-1} ip(i)$

Overview

Introduction

- Problem
- Objective

Alignment-free methods base on image textures

- First-order statistics (FOS)
- Gray Level Co-ocurrence Matrix (GLCM)
- Multi-resolution Local Binary Patterns (MLBP)

Results

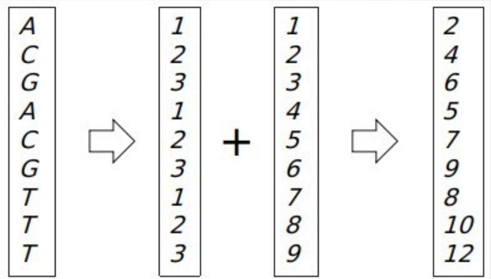
- Datasets
- Comparison of FOS, GLCM, LBP and MLBP
- Comparison of mapping functions

Conclusions

GLCM

Mapping function

Each base in sequence $S = \{A, C, G, T\}$ is mapped to the numbers $S' = \{1, 2, 3, 4\}$. Then we added to each value the base position.

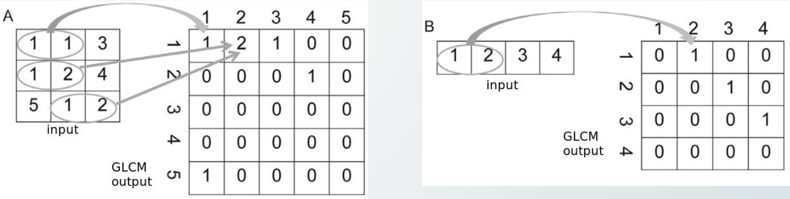


A	⇨	1	+	1	⇨	2
C		2		2		4
G		3		3		6
A		1		4		5
C		2		5		7
G		3		6		9
T		1		7		8
T		2		8		10
T		3		9		12

Then, compute gray-level co-occurrence matrix.

GLCM

Mapping function



1	1	3
1	2	4
5	1	2

	1	2	3	4	5
1	1	2	1	0	0
2	0	0	0	1	0
3	0	0	0	0	0
4	0	0	0	0	0
5	1	0	0	0	0

1	2	3	4

	1	2	3	4
1	0	1	0	0
2	0	0	1	0
3	0	0	0	1
4	0	0	0	0

Figure: Examples of GLCM algorithm. Left: GLCM computed from a 2D matrix with intensities from 1 to 5. Right: GLCM computed from a 1D vector with intensities from 1 to 4.

GLCM
Features

From the histogram, the following features are compute:

- ***Entropy*** $= -\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)Ln(p(i,j))$
- ***Contrast*** $= \sum_{i=1}^{L}\sum_{j=1}^{L}(i-j)^2 p(i,j)$
- ***Energy*** $= \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)^2$
- ***Correlation*** $= \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{(i-\mu_i)(j-\mu_i)p(i,j)}{\sigma_i \sigma_j}$
- ***Homogeneity*** $= \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{p(i,j)}{1+|i-j|}$

where, $p(i,j)$ is the GLCM matrix and L is the maximun intensity value.

Overview

Introduction

- Problem
- Objective

Alignment-free methods base on image textures

- First-order statistics (FOS)
- Gray Level Co-ocurrence Matrix (GLCM)
- Multi-resolution Local Binary Patterns (MLBP)

Results

- Datasets
- Comparison of FOS, GLCM, LBP and MLBP
- Comparison of mapping functions

Conclusions

MLBP

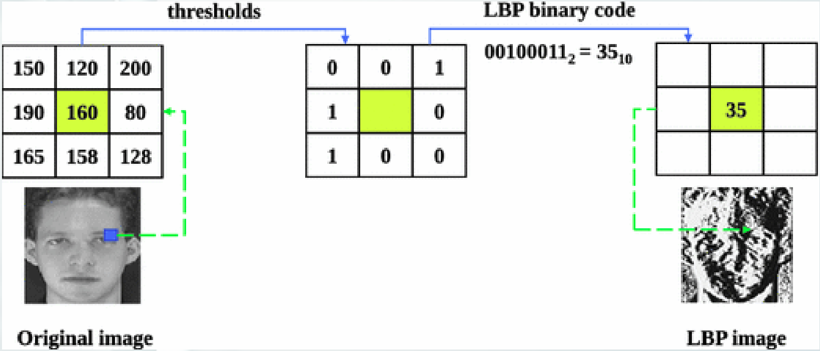
Mapping function

Table: Numeric representation for each base used by Kouchaki et al. [7].

Base	Integer	EIIP	Atomic	Real
A	2	0.1260	70	-1.5
T	-2	0.1335	78	1.5
C	-1	0.1340	58	-0.5
G	2	0.0806	66	0.5

MLBP

LBP example



thresholds

150	120	200
190	160	80
165	158	128

Original image

0	0	1
1		0
1	0	0

LBP binary code

$00100011_2 = 35_{10}$

	35	

LBP image

Figure: Example of Local Binary Pattern algorithm.

MLBP
Features

$$LBP(x(t)) = \sum_{i=0}^{p/2-1} (Sign(x(t+i-p/2)-x(t))2^{i}+ \\ Sign(x(t+i+1)-x(t))2^{i+p/2}), \tag{2}$$

where p in the number of neighbouring points and *Sign* is:

$$Sign(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{3}$$

$$h_k = \sum_{p/2 \leq i \leq N-p/2} \delta(LBP_p(x(i), k)), \tag{4}$$

Overview

Introduction

- Problem
- Objective

Alignment-free methods base on image textures

- First-order statistics (FOS)
- Gray Level Co-ocurrence Matrix (GLCM)
- Multi-resolution Local Binary Patterns (MLBP)

Results

- Datasets
- Comparison of FOS, GLCM, LBP and MLBP
- Comparison of mapping functions

Conclusions

Datasets

Table: 16S ribosomal DNA of 13 bacteria.

Species	Accesion Code	Length (bp)
Bacillus maritimus	KP317497	1515
Bacillus wakoensis	NR_040849	1524
Bacillus australimaris	NR_148787	1513
Bacillus xiamenensis	NR_148244	1513
Escherichia coli	J01859	1541
Streptococcus himalayensis	NR_156072	1509
Streptococcus halotolerans	NR_152063	1520
Streptococcus tangierensis	NR_134818	1520
Streptococcus cameli	NR_134817	1518
Thermus amyloliquefaciens	NR_136784	1514
Thermus tengchongensis	NR_132306	1523
Thermus thermophilus	NR_037066	1515
Thermus filiformis	NR_117152	1514

Datasets

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

Species	Accesion Code	Length (bp)
Macaca fascicularis	M22653	896
Macaca fuscata	M22651	896
Macaca mulatta	M22650	896
Macaca sylvanus	M22654	896
Saimiri sciureus	M22655	893
Chimpanzee	V00672	896
Lemur catta	M22657	895
Gorilla	V00658	896
Hylobates	V00659	896
Sumatran Orangutan	V00675	895
Tarsius syrichta	M22656	895
Human	L00016	896

Datasets

Table: The mitochondrial genome detailed information of 18 eutherian mammals from NCBI database.

Species	Accesion Code	Length (bp)
Human	V00662	16569
Pygmy chimpanzee	D38116	16563
Common chimpanzee	D38113	16554
Gorilla	D38114	16364
Orangutan	D38115	16389
Gibbon	X99256	16472
Baboon	Y18001	16521
Horse	X79547	16660
White rhinoceros	Y07726	16832
Harbor seal	X63726	16826
Gray seal	X72004	16797
Cat	U20753	17009
Fin whale	X61145	16397
Blue whale	X72204	16402
Cow	V00654	16338

Mapping functions

Table: Numeric representation for each base.

Base	MAP0	MAP1	MAP2	MAP3	MAP4	MAP5
A	FOS’s prop.	GLCM’s prop.	2	0.1260	70	-1.5
T	FOS’s prop.	GLCM’s prop.	-2	0.1335	78	1.5
C	FOS’s prop.	GLCM’s prop.	-1	0.1340	58	-0.5
G	FOS’s prop.	GLCM’s prop.	2	0.0806	66	0.5

Overview

Introduction

- Problem
- Objective

Alignment-free methods base on image textures

- First-order statistics (FOS)
- Gray Level Co-ocurrence Matrix (GLCM)
- Multi-resolution Local Binary Patterns (MLBP)

Results

- Datasets
- Comparison of FOS, GLCM, LBP and MLBP
- Comparison of mapping functions

Conclusions

Results

Comparison in 16S ribosomal DNA dataset

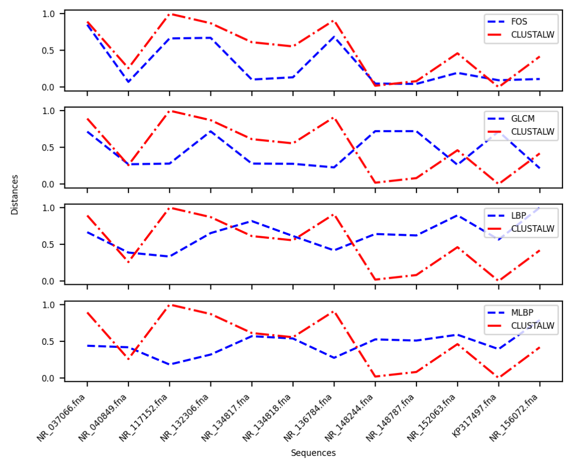


Figure: Euclidean distance of Escherichia coli against the rest sequences in 16S ribosomal DNA dataset. We used MEGA. FOS, GLCM, LBP and MLBP.

Results

Comparison in NADH dataset

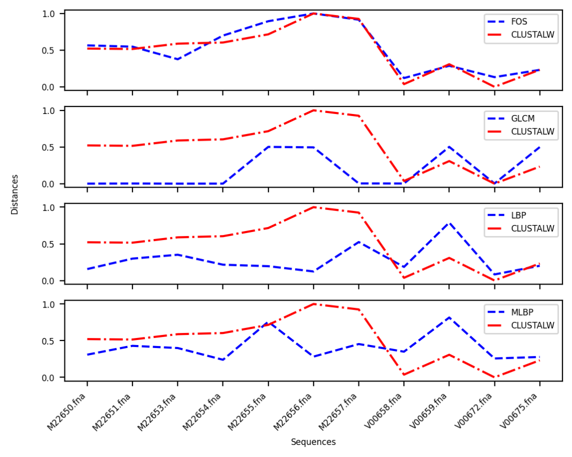


Figure: Euclidean distance of Human against the rest sequences in NADH dehydrogenase protein dataset. We used MEGA. FOS, GLCM, LBP and MLBP.

Results

Comparison in the mitochondrial dataset

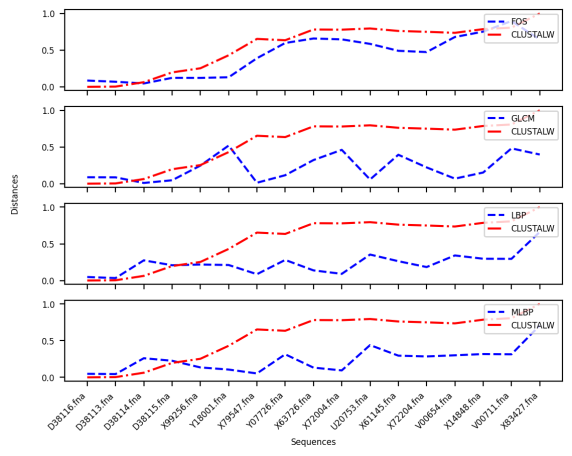


Figure: Euclidean distance of Human against the rest sequences in the mitochondrial genome dataset. We used MEGA. FOS, GLCM, LBP and MLBP.

Overview

Introduction

- Problem
- Objective

Alignment-free methods base on image textures

- First-order statistics (FOS)
- Gray Level Co-ocurrence Matrix (GLCM)
- Multi-resolution Local Binary Patterns (MLBP)

Results

- Datasets
- Comparison of FOS, GLCM, LBP and MLBP
- Comparison of mapping functions

Conclusions

Results

Comparison of the six mapping functions using FOS algorithm

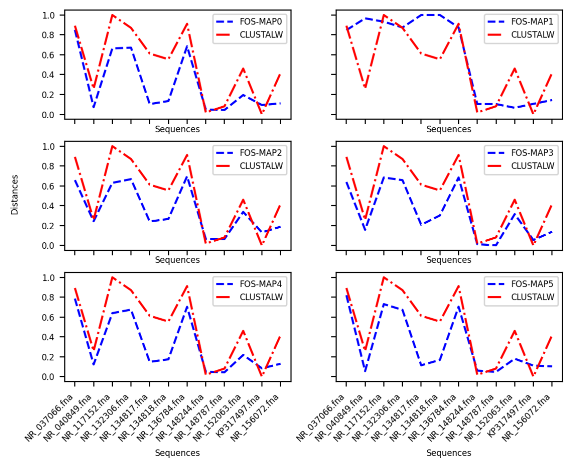


Figure: Comparison of the 6 mapping functions using FOS algorithm over the 16S ribosomal DNA dataset.

Results

Comparison of MAP1 mapping function

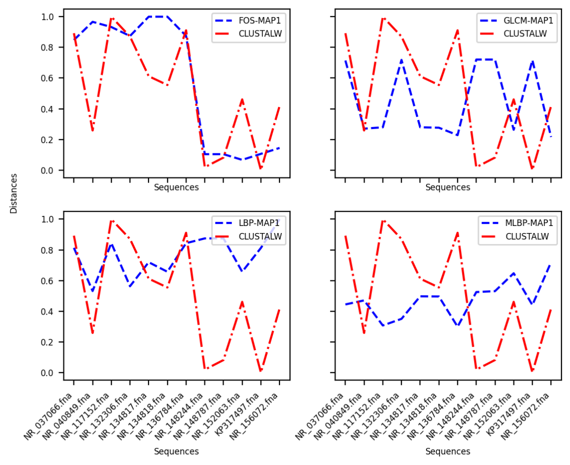


Figure: Comparison of MAP1 mapping function over the 16S ribosomal DNA dataset.

Results

Comparison of the six mapping functions using GLCM algorithm

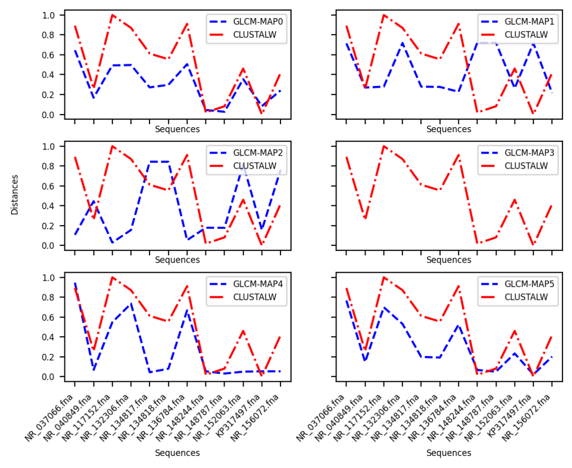
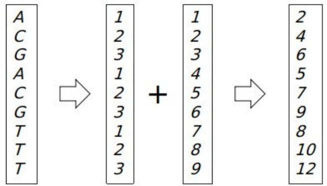


Figure: Comparison of the 6 mapping functions using GLCM algorithm over the 16S ribosomal DNA dataset.

Results

Discussion: Problem with MAP1 function

MAP1 function, proposed by Chen at el. [6].



A		1		1		2
C		2		2		4
G		3		3		6
A		1		4		5
C	⇨	2	+	5	⇨	7
G		3		6		9
T		1		7		8
T		2		8		10
T		3		9		12

The resultant vector have disperse values and it depends strongly from the sequence's length

Results

Comparison of the six mapping functions using LBP algorithm

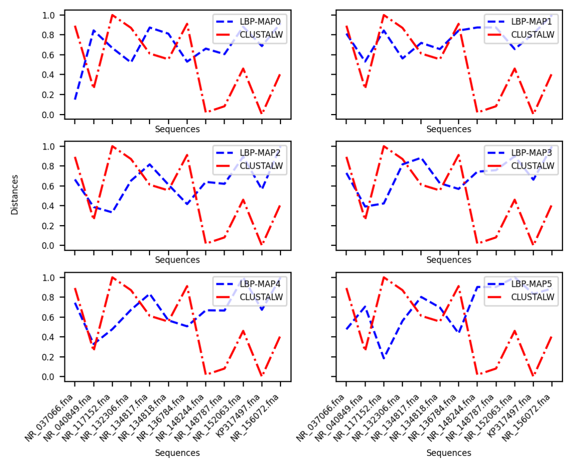


Figure: Comparison of the 6 mapping functions using the LBP algorithm over the 16S ribosomal DNA dataset.

Results

Comparison of the six mapping functions using MLBP algorithm

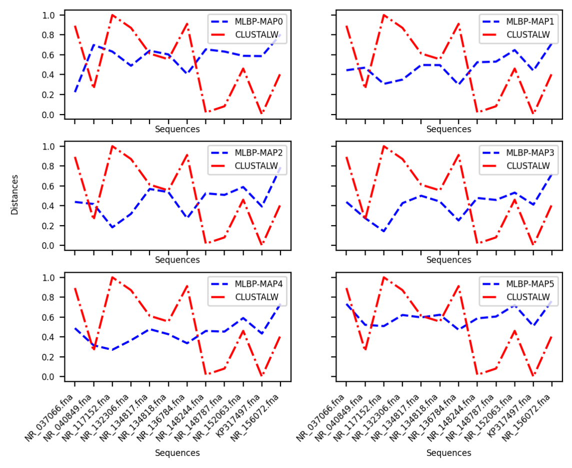


Figure: Comparison of the 6 mapping functions using the MLBP algorithm over the 16S ribosomal DNA dataset.

Results

Square error of mapping functions and algorithms over the 16S ribosomal DNA dataset

Table: Square error of all mapping functions and the four algorithms over the 16S ribosomal DNA dataset.

Mapping function	FOS	GLCM	LBP	MLBP
MAP0	0.07093	0.07214	0.2498	0.1997
MAP1	0.09229	0.22709	0.2187	0.1805
MAP2	**0.04875**	0.27343	0.1977	0.1965
MAP3	0.05038		0.2106	0.1848
MAP4	0.06267	0.09814	0.1997	0.1630
MAP5	0.06592	0.06572	0.3395	0.1369

Results

Square error of mapping functions and algorithms over the NADH dataset

Table: Square error of all mapping functions and the four algorithms over the NADH dataset.

Mapping function	FOS	GLCM	LBP	MLBP
MAP0	**0.0103**	0.0345	0.0406	0.0711
MAP1	0.1795	0.2279	0.2029	0.0895
MAP2	0.0126	0.0307	0.1682	0.1258
MAP3	0.1642		0.1310	0.1022
MAP4	0.0297	0.0784	0.1410	0.0630
MAP5	0.0865	0.0345	0.0792	0.0452

Results

Square error of mapping functions and algorithms over the mitochondrial genome dataset

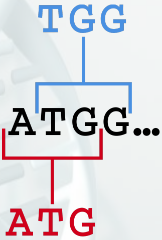
Table: Square error of all mapping functions and the four algorithms over the mitochondrial genome dataset.

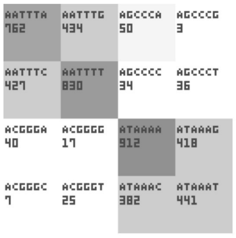
Mapping function	FOS	GLCM	LBP	MLBP
MAP0	**0.0329**	0.0569	0.1693	0.1254
MAP1	0.2439	0.1951	0.1294	0.1465
MAP2	0.0417	0.1567	0.1746	0.1654
MAP3	0.1811		0.0768	0.1094
MAP4	0.0570	0.0731	0.1765	0.1724
MAP5	0.1255	0.0851	0.1622	0.1575

Results

Discussion: Best mapping function

MAP0 function and histogram proposed by Deliba et al. [3] is very similar to k-mer frecuencies. [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]





AATTTA 762	AATTTG 434	AGCCCA 50	AGCCCG 3
AATTTC 427	AATTTT 830	AGCCCC 34	AGCCCT 36
ACGGGA 40	ACGGGG 17	ATAAAA 912	ATAAAG 418
ACGGGC 7	ACGGGT 25	ATAAAC 382	ATAAAT 441

LBP and MLBP reflects the correlation among pixels within a local area, but the main infomration in DNA sequences is no related to correlations of neighbors bases.

Results

Robinson Fould distances of phylogenetic trees

Table: Phylogenetics tree distances using Robinson Fould algorithm.

Database	FOS	GLCM	LBP	MLBP
16S ribosomal	14/20	18/20	18/20	**12/20**
NADH	**12/18**	18/18	18/18	16/18
Mitochondrial	**14/30**	30/30	30/30	16/30

Results

Phylogenetic trees got by Phylo.io

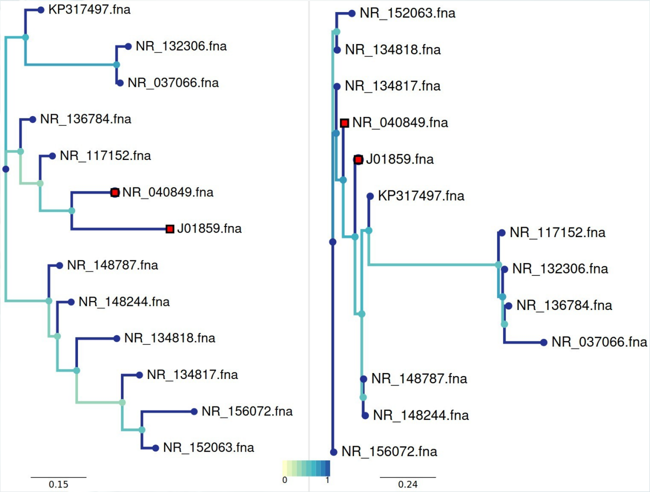


Figure: Phylogenetics tree of MEGA and FOS in the 16S ribosomal dataset.

Conclusions

- We compared FOS, GLCM, LBP, and MLBP with six mapping functions. We also, compare the phylogenetic trees with Robinson Fould algorithm and Phylo.io.

Conclusions

- We compared FOS, GLCM, LBP, and MLBP with six mapping functions. We also, compare the phylogenetic trees with Robinson Fould algorithm and Phylo.io.
- FOS got the best results. Moreover, MAP1 was the worst mapping function and MAP0 was the best because of its similarity to k-mer method.

Conclusions

- We compared FOS, GLCM, LBP, and MLBP with six mapping functions. We also, compare the phylogenetic trees with Robinson Fould algorithm and Phylo.io.
- FOS got the best results. Moreover, MAP1 was the worst mapping function and MAP0 was the best because of its similarity to k-mer method.
- LBP and MLBP are not suitable for sequence similarity because they consider the correlation between neighbors.

Conclusions

- We compared FOS, GLCM, LBP, and MLBP with six mapping functions. We also, compare the phylogenetic trees with Robinson Fould algorithm and Phylo.io.
- FOS got the best results. Moreover, MAP1 was the worst mapping function and MAP0 was the best because of its similarity to k-mer method.
- LBP and MLBP are not suitable for sequence similarity because they consider the correlation between neighbors.
- Furthermore, FOS's tree is the most similar to MEGA's tree for the NADH dehydrogenase and the mitochondrial genomes datasets.

References I

[1] J. M. Archibald, *Genomics: A Very Short Introduction*. Oxford University Press, 2018, vol. 559.

[2] G. S. Randhawa, M. P. Soltysiak, H. El Roz, C. P. de Souza, K. A. Hill, and L. Kari, “Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study,” *Plos one*, vol. 15, no. 4, p. e0232391, 2020.

[3] E. Delibaş and A. Arslan, “Dna sequence similarity analysis using image texture analysis based on first-order statistics,” *Journal of Molecular Graphics and Modelling*, p. 107603, 2020.

[4] X. Jin, Q. Jiang, Y. Chen, S.-J. Lee, R. Nie, S. Yao, D. Zhou, and K. He, “Similarity/dissimilarity calculation methods of dna sequences: a survey,” *Journal of Molecular Graphics and Modelling*, vol. 76, pp. 342–355, 2017.

References II

[5] J. Xiong, *Essential bioinformatics*. Cambridge University Press, 2006.

[6] W. Chen, B. Liao, and W. Li, "Use of image texture analysis to find dna sequence similarities," *Journal of theoretical biology*, vol. 455, pp. 1–6, 2018.

[7] S. Kouchaki, A. Tapinos, and D. L. Robertson, "A signal processing method for alignment-free metagenomic binning: Multi-resolution genomic binary patterns," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[8] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.

References III

[9] O. Robinson, D. Dylus, and C. Dessimoz, “Phylo. io: interactive viewing and comparison of large phylogenetic trees on the web,” *Molecular biology and evolution*, vol. 33, no. 8, pp. 2163–2166, 2016.

[10] S. Karlin and I. Ladunga, “Comparisons of eukaryotic genomic sequences,” *Proceedings of the National Academy of Sciences*, vol. 91, no. 26, pp. 12 832–12 836, 1994.

[11] A. Campbell, J. Mrazek, and S. Karlin, “Genome signature comparisons among prokaryote, plasmid, and mitochondrial dna,” *Proceedings of the National Academy of Sciences*, vol. 96, no. 16, pp. 9184–9189, 1999.

References IV

[12] A. M. Shedlock, C. W. Botka, S. Zhao, J. Shetty, T. Zhang, J. S. Liu, P. J. Deschavanne, and S. V. Edwards, “Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 8, pp. 2767–2772, 2007.

[13] T.-J. Wu, Y.-C. Hsieh, and L.-A. Li, “Statistical measures of dna sequence dissimilarity under markov chain models of base composition,” *Biometrics*, vol. 57, no. 2, pp. 441–448, 2001.

[14] G. E. Sims, S.-R. Jun, G. A. Wu, and S.-H. Kim, “Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 8, pp. 2677–2682, 2009.

References V

[15] G. E. Sims and S.-H. Kim, “Whole-genome phylogeny of escherichia coli/shigella group by feature frequency profiles (ffps),” *Proceedings of the National Academy of Sciences*, vol. 108, no. 20, pp. 8329–8334, 2011.

[16] T.-J. Wu, Y.-H. Huang, and L.-A. Li, “Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between dna sequences,” *Bioinformatics*, vol. 21, no. 22, pp. 4125–4132, 2005.

[17] Q. Dai, Y. Yang, and T. Wang, “Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison,” *Bioinformatics*, vol. 24, no. 20, pp. 2296–2302, 2008.

[18] B. Haubold, “Alignment-free phylogenetics and population genetics,” *Briefings in bioinformatics*, vol. 15, no. 3, pp. 407–418, 2014.

References VI

[19] R. Karamichalis, L. Kari, S. Konstantinidis, and S. Kopecki, “An investigation into inter-and intragenomic variations of graphic genomic signatures,” *BMC bioinformatics*, vol. 16, no. 1, p. 246, 2015.

[20] S. Vinga and J. Almeida, “Alignment-free sequence comparison—a review,” *Bioinformatics*, vol. 19, no. 4, pp. 513–523, 2003.