UNIVERSIDAD NACIONAL DE SAN AGUSTÍN

Mineria de Datos

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

MSc. Vicente Machaca Arceda

May 22, 2020

Overview

Introduction

Problem

Solution

Results

Conclusion

Overview

Introduction

Problem

Solution

Results

Conclusion

Introduction
DNA

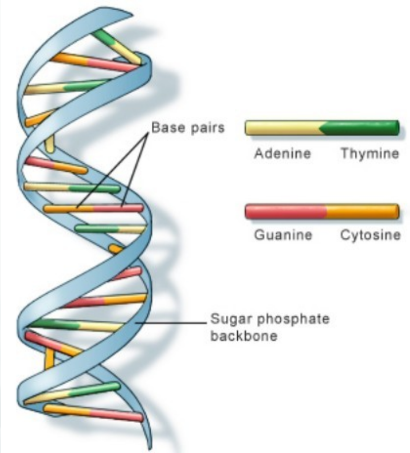


Figure: Molecules in DNA. Adenine, Thymine, Guanine and Cytosine [1].

Introduction

DNA

The human genome is made of ~**3.2 billions bp** of DNA.
~6.4 billions of nucleotides [2].

The HIV-1 genome is made of ~**20k bp** of DNA.
Meanwhile, the COVID-19 is made of ~**32k bp** [3].

Table: Total GigaBytes used to store a complete diploid genome.

DNA bases	4
Bits per base	2
Base pairs per genome	3200000000
Bits per genome	12800000000
Total bits Diplod genome	25600000000
Total Kilobytes	3125000
Total Megabytes	3052
Total Gigabytes	**2.980**

Introduction

Visualization

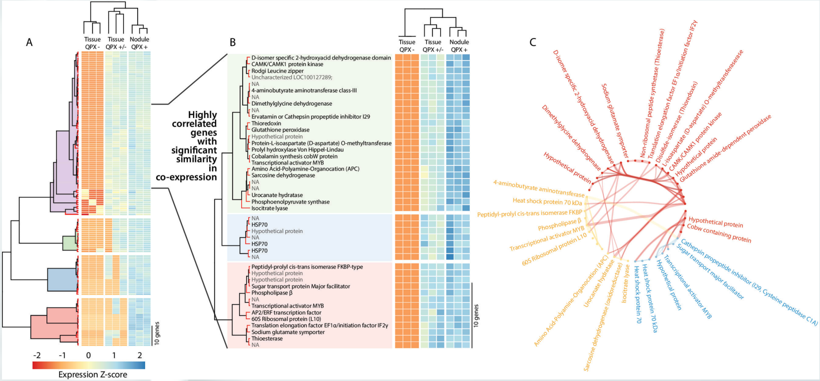


Figure: Example of visualization in bioinformatics.

Similarity

It is used for identifying evolutionary or affinity relations [4]. Similarity analysis is an important research area in Bioinformatics [5].

Phylogenetics

Phylogenetics is the study of the evolutionary history of living organisms using tree- like diagrams to represent pedigrees of these organisms [6].

Introduction
Phylogenetics

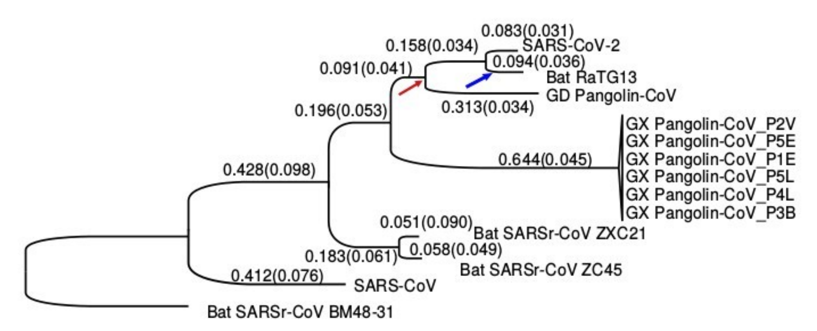


Figure: The phylogenetic tree of SARS-CoV-2 (COVID-19) and the related Coronaviruses [7].

Introduction

Phylogenetics

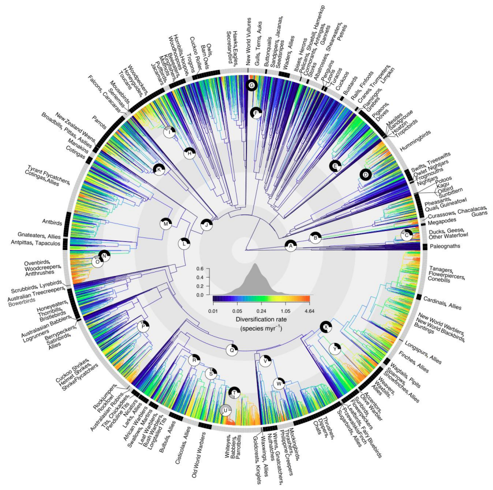
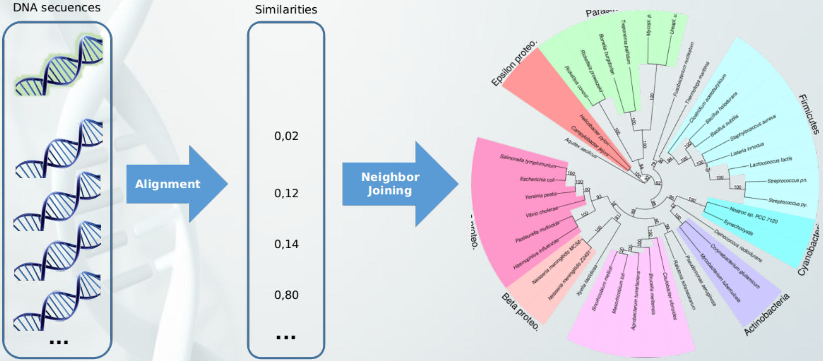

Paleognaths
Ducks, Geese, Other Waterfowl
Megapodes
Curassows, Chacalacas, Guans
Pheasants, Quail, Guineafowl
Potoos
Kagu
Oilbird
Sunbittern
Swifts, Treeswifts
Owlet Nightjars
Frogmouths
Nightjars
Hummingbirds
Mesites
Sandgrouse
Hoatzin
Tropicbirds
Pigeons, Doves
Flamingos, Grebes
Cranes, Trumpeters, Limpkin
Rails, Finfoots
Bustards
Cuckoos
Turacos
Loons
Penguins
Albatrosses, Shearwaters, Petrels
Storks
Cormorants, Anhingas, Gannets
Pelicans, Shoebill, Hamerkop
Ibises, Herons
Waders, Allies
Sandpipers, Jacanas, Seedsnipes
Buttonquails
Gulls, Terns, Auks
New World Vultures
Hawks,Eagles, Secretarybird
Owls, Barn Owls
Cuckoo Roller
Trogons
Hornbills,Hoopoe, Woodhoopoes
Beeeaters, Kingfishers, Motmots
Puffbirds, Jacamars
Woodpeckers, Honeyguides, Toucans
Mousebirds
Seriemas
Falcons, Caracaras
Parrots
New Zealand Wrens
Broadbills, Pittas, Asities
Manakins
Cotingas
Tyrant Flycatchers, Cotingas,Allies
Antbirds
Gnateaters, Allies
Antpittas, Tapaculos
Ovenbirds, Woodcreepers, Antthrushes
Scrubbirds, Lyrebirds
Australian Treecreepers, Bowerbirds
Honeyeaters, Thornbills, Bristlebirds
Australasian Babblers, Logrunners
Berrypeckers, Satinbirds, Allies
Cuckoo Shrikes, Helmet Shrikes, ShrikeFlycatchers
Rockjumpers, Rockfowl
Australasian Robins
Tits, Chickadees, Penduline Tits
Nicators
Larks, Allies
African Warblers
Swallows, Martins
Leaf Warblers, Bush Warblers, Longtailed Tits
Bulbuls, Allies
Cisticolas, Allies
Old World Warblers
Whiteyes, Babblers, Parrotbills
Goldcrests, Kinglets
Waxwings, Allies
Wrens, Gnatcatchers, Nuthatches
Mockingbirds, Thrashers, Philippine Creepers
Thrushes, Dippers, Chats
Sugarbirds, Allies
PrzewalskisFinch
Leafbirds, Fairy Bluebirds
Sunbirds, Flowerpeckers
Accentors, Olive Warbler
Weavers, Waxbills, Allies
Sparrows, Snowfinches, Allies
Wagtails, Pipits
Finches, Allies
Longspurs, Allies
New World Warblers, New World Blackbirds, Buntings
Cardinals, Allies
Tanagers, Flowerpiercers, Conebills
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
0.6
0.4
0.2
0.0
0.01
0.05
0.24
1.05
4.64
Diversification rate
(species myr^{-1})

Figure: The phylogeny tree of bird species [8].

Introduction

Phylogenetics steps



DNA secuences

...

Alignment

Similarities

0,02

0,12

0,14

0,80

...

Neighbor Joining

Figure: Steps to visualize phylonetics trees.

Overview

Introduction

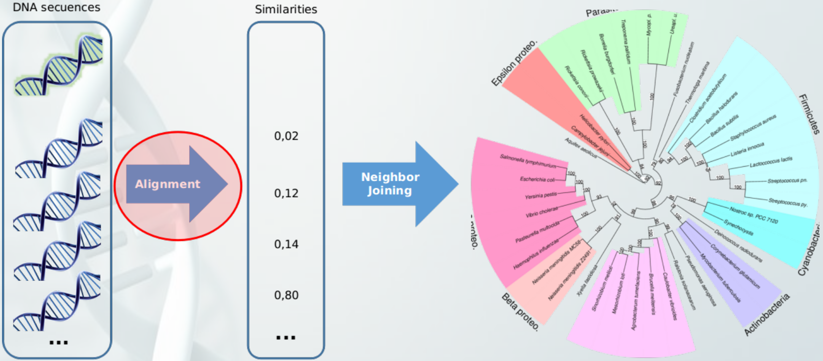
Problem

Solution

Results

Conclusion

Problem



DNA secuences

...

Alignment

Similarities

0,02

0,12

0,14

0,80

...

Neighbor Joining

Figure: Steps to visualize phylonetics trees.

Problem

Alignment-based methods

- The most used **alignment-based** method is BLAST.

Problem

Alignment-based methods

- The most used **alignment-based** method is BLAST.
- BLAST is slow.

Problem

Alignment-based methods

- The most used **alignment-based** method is BLAST.
- BLAST is slow.
- DNA sequences increases every day so BLAST get slower every second.

Problem

Alignment-based methods

- The most used **alignment-based** method is BLAST.
- BLAST is slow.
- DNA sequences increases every day so BLAST get slower every second.
- For example, two days were necessary to build the Phylogenetis tree of COVID-19.

Overview

Introduction

Problem

Solution

Results

Conclusion

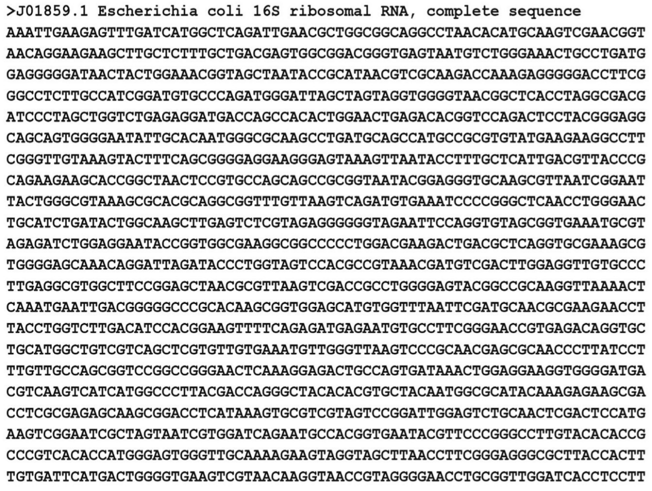
Solution

Alignment-free methods

There are **alignment-free** methods that convert the DNA sequences into feature vectors, then it is easy two compute distances (similarities) [4].

Solution

Image texture analysis



```
>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.

Solution

Image texture analysis

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

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

Solution

Image texture analysis

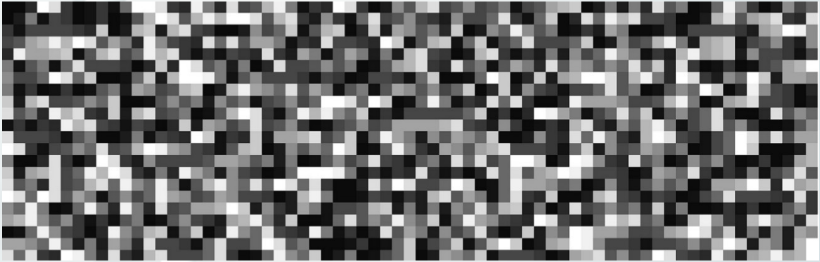


Figure: Textures converted from the DNA sequences. Source [4].

Solution

Image texture analysis

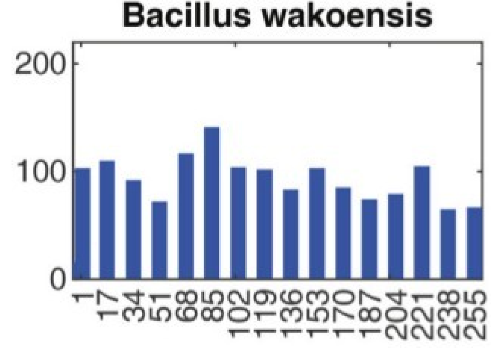


Figure: Histogram of 16S ribosomal DNA. Source [4].

Solution

Image texture analysis

From the histogram, the following features are compute:

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

Solution

Results

Conclusion

Results

Dataset

Table: 16S ribosomal DNA of 13 bacteria

Specie	Accession 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 thermophilus	NR_117152	1514

Results

Phylogenetic tree

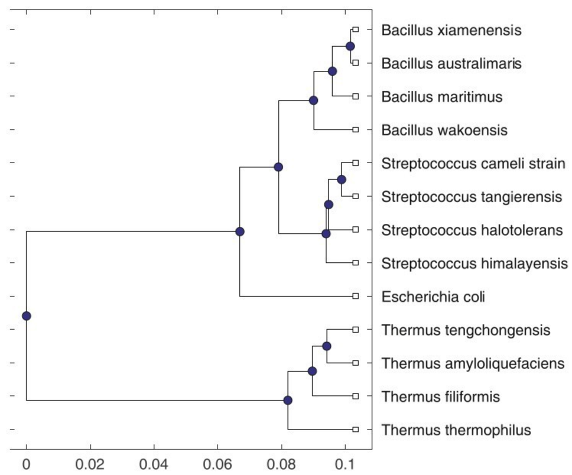


Figure: Phylogenetic tree generated by the proposed method. Source: [4]

Results

Phylogenetic tree

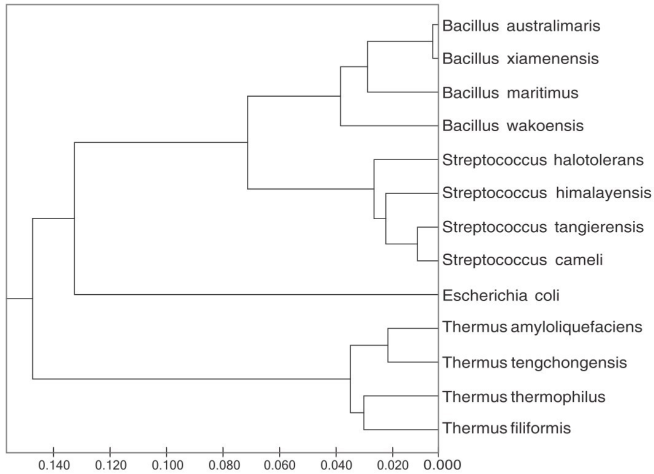


Figure: Phylogenetic tree generated by MEGA7 based on ClustalW alignment and the UPGMA method. Source: [4]

Results

Phylogenetic tree

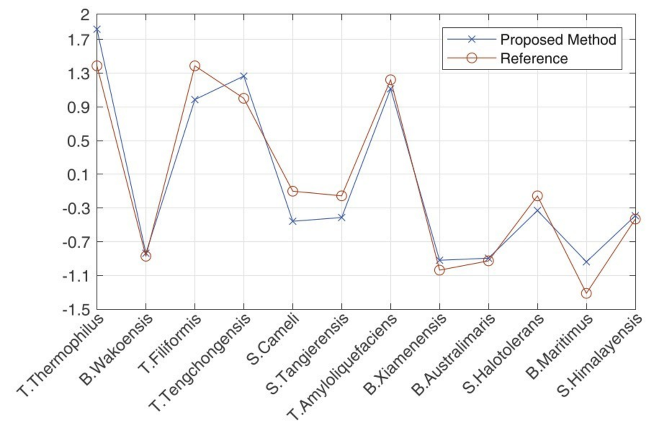


Figure: The degree of similarity/dissimilarity of the other 12 bacteria and Escherichia coli. Source: [4]

Overview

Introduction

Problem

Solution

Results

Conclusion

Conclusion

- An image texture from DNA is proposed for DNA analysis similarity.
- The method proposed results in a phylogenetic tree very similar to the result of MEGA.
- The method proposed have a low time processing but the authors did not measure it.
- Moreover, It is necessary an evaluation with complete genomes and more samples.

References I

[1] M. Clinics, “How genetic disorders are inherited,” https://www.mayoclinic.org/tests-procedures/genetic-testing/multimedia/genetic-disorders/sls-20076216?s=2, 2020, accessed: 2020-03-20.

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

[3] 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,” *bioRxiv*, 2020.

[4] 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.

References II

[5] 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.

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

[7] X. Tang, C. Wu, X. Li, Y. Song, X. Yao, X. Wu, Y. Duan, H. Zhang, Y. Wang, Z. Qian *et al.*, “On the origin and continuing evolution of sars-cov-2,” *National Science Review*, 2020.

[8] P. Mannion, “First ever family tree for all living birds reveals evolution and diversification,” 2020.