Árboles Filogéneticos

Daniel Barreras, Yair Beltrán, Victor Symonds

2023-05-01

Instala los paquetes y carga las librerías necesarias para hacer un análisis filogenético de genomas virales en R. Por ejemplo, ejecuta los siguientes comandos una vez que hayas instalado los paquetes necesarios.

library(viridis)

## Warning: package 'viridis' was built under R version 4.2.3

## Loading required package: viridisLite

library(Biostrings)

## Loading required package: BiocGenerics

##   
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':  
##   
## IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':  
##   
## anyDuplicated, aperm, append, as.data.frame, basename, cbind,  
## colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,  
## get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,  
## match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,  
## Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,  
## table, tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##   
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':  
##   
## expand.grid, I, unname

## Loading required package: IRanges

##   
## Attaching package: 'IRanges'

## The following object is masked from 'package:grDevices':  
##   
## windows

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##   
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':  
##   
## strsplit

library(DECIPHER)

## Loading required package: RSQLite

## Warning: package 'RSQLite' was built under R version 4.2.3

## Loading required package: parallel

library(ade4)

## Warning: package 'ade4' was built under R version 4.2.3

##   
## Attaching package: 'ade4'

## The following object is masked from 'package:Biostrings':  
##   
## score

## The following object is masked from 'package:BiocGenerics':  
##   
## score

library(seqinr)

## Warning: package 'seqinr' was built under R version 4.2.3

##   
## Attaching package: 'seqinr'

## The following object is masked from 'package:Biostrings':  
##   
## translate

library(adegenet)

## Warning: package 'adegenet' was built under R version 4.2.3

##   
## /// adegenet 2.1.10 is loaded ////////////  
##   
## > overview: '?adegenet'  
## > tutorials/doc/questions: 'adegenetWeb()'   
## > bug reports/feature requests: adegenetIssues()

library(ape)

## Warning: package 'ape' was built under R version 4.2.3

##   
## Attaching package: 'ape'

## The following objects are masked from 'package:seqinr':  
##   
## as.alignment, consensus

## The following object is masked from 'package:Biostrings':  
##   
## complement

library(ggtree)

## ggtree v3.6.2 For help: https://yulab-smu.top/treedata-book/  
##   
## If you use the ggtree package suite in published research, please cite  
## the appropriate paper(s):  
##   
## Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam.  
## ggtree: an R package for visualization and annotation of phylogenetic  
## trees with their covariates and other associated data. Methods in  
## Ecology and Evolution. 2017, 8(1):28-36. doi:10.1111/2041-210X.12628  
##   
## LG Wang, TTY Lam, S Xu, Z Dai, L Zhou, T Feng, P Guo, CW Dunn, BR  
## Jones, T Bradley, H Zhu, Y Guan, Y Jiang, G Yu. treeio: an R package  
## for phylogenetic tree input and output with richly annotated and  
## associated data. Molecular Biology and Evolution. 2020, 37(2):599-603.  
## doi: 10.1093/molbev/msz240  
##   
## Shuangbin Xu, Lin Li, Xiao Luo, Meijun Chen, Wenli Tang, Li Zhan, Zehan  
## Dai, Tommy T. Lam, Yi Guan, Guangchuang Yu. Ggtree: A serialized data  
## object for visualization of a phylogenetic tree and annotation data.  
## iMeta 2022, 4(1):e56. doi:10.1002/imt2.56

##   
## Attaching package: 'ggtree'

## The following object is masked from 'package:ape':  
##   
## rotate

## The following object is masked from 'package:Biostrings':  
##   
## collapse

## The following object is masked from 'package:IRanges':  
##   
## collapse

## The following object is masked from 'package:S4Vectors':  
##   
## expand

library(viridis)  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

Obtén las secuencias de 10 genomas virales a partir del código de acceso al genoma en GenBank usando la función read.GenBank del paquete ape y muestra el código empleado para hacerlo.

coronavirus <- c("MT810758", "AY278489", "MW030193", "AY508724", "MT594401",  
 "AY485277", "MT292571", "AY390556", "MN908947","JX869059",  
 "MW133981", "MT577009", "MT835383", "MW056032", "MT470219")  
  
  
coronavirus <- c("MW133981", "MT577009" ,"MT835383", "MT890462", "MW056032",  
 "MT470219", "MT594401", "MW030193", "MT810758", "MW041156",   
 "MT324062", "MT994849", "MT670013", "MT940481", "MW134558",  
 "MT876433", "MT845878", "MT919790", "MT955161","MT820489")  
  
coronavirus\_sequences <- read.GenBank(coronavirus)

Imprime en consola la estructura del objeto de clase DNABin que obtuviste con la función read.GenBank. Muestra el código empleado para hacerlo y la estructura del objeto DNAbin que obtuviste.

str(coronavirus\_sequences)

## List of 20  
## $ MW133981: raw [1:29866] 88 88 28 18 ...  
## $ MT577009: raw [1:29903] 88 18 18 88 ...  
## $ MT835383: raw [1:29858] 88 18 48 18 ...  
## $ MT890462: raw [1:29903] 88 18 18 88 ...  
## $ MW056032: raw [1:29826] 88 88 28 88 ...  
## $ MT470219: raw [1:29903] 88 18 18 88 ...  
## $ MT594401: raw [1:29804] 88 88 28 18 ...  
## $ MW030193: raw [1:29520] 28 18 88 18 ...  
## $ MT810758: raw [1:29409] 88 18 48 48 ...  
## $ MW041156: raw [1:29870] 88 18 18 88 ...  
## $ MT324062: raw [1:29903] 88 18 18 88 ...  
## $ MT994849: raw [1:29819] 48 88 18 28 ...  
## $ MT670013: raw [1:29782] 88 48 88 18 ...  
## $ MT940481: raw [1:29821] 88 28 18 18 ...  
## $ MW134558: raw [1:29806] 88 28 28 88 ...  
## $ MT876433: raw [1:29903] 88 18 18 88 ...  
## $ MT845878: raw [1:29835] 88 88 28 88 ...  
## $ MT919790: raw [1:29815] 88 28 18 18 ...  
## $ MT955161: raw [1:29852] 88 18 18 88 ...  
## $ MT820489: raw [1:29825] 88 18 88 28 ...  
## - attr(\*, "class")= chr "DNAbin"  
## - attr(\*, "description")= chr [1:20] "MW133981.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/TX-DSHS-0512/2020, complete genome" "MT577009.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/IND/763/2020, complete genome" "MT835383.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/BRA/RJ-DCVN5/2020, complete genome" "MT890462.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/RUS/20200417\_10/2020, complete genome" ...  
## - attr(\*, "species")= chr [1:20] "Severe\_acute\_respiratory\_syndrome\_coronavirus\_2" "Severe\_acute\_respiratory\_syndrome\_coronavirus\_2" "Severe\_acute\_respiratory\_syndrome\_coronavirus\_2" "Severe\_acute\_respiratory\_syndrome\_coronavirus\_2" ...

Concentra en un archivo todas las secuencias de los genomas con la función write.dna del paquete ape y muestra el código empleado para hacerlo.

write.dna(coronavirus\_sequences, file = "coronavirus\_sequences.fasta", format = "fasta")

Carga las secuencias concentradas en el archivo del punto anterior con la función readDNAStringSet de Biostrings. Muestra el código empleado para hacerlo, imprime en consola el contenido del objeto tipo DNAStringSet e inclúyelo en tu entregable.

my\_dna\_string\_set <- readDNAStringSet("coronavirus\_sequences.fasta", format = "fasta")

## Warning in .Call2("fasta\_index", filexp\_list, nrec, skip, seek.first.rec, :  
## reading FASTA file coronavirus\_sequences.fasta: ignored 49685 invalid  
## one-letter sequence codes

my\_dna\_string\_set

## DNAStringSet object of length 20:  
## width seq names   
## [1] 29866 AACTTTCGATCTCTTGTAGATCT...AAAAAAAAAAAAAAAAAAAAAA MW133981  
## [2] 29903 ATTAAAGGTTTATACCTTCCCAG...AAAAAAAAAAAAAAAAAAAAAA MT577009  
## [3] 29858 ATGTTTATACCTTCCCAGGTAAC...TGATTTTAATAGCTTCTTAGGA MT835383  
## [4] 29903 ATTAAAGGTTTATACCTTCCCAG...AAAAAAAAAAAAAAAAAAAAAA MT890462  
## [5] 29826 AACAAACCAACCAACTTTCGATC...GCTATCCCCATGTGATTTTAAT MW056032  
## ... ... ...  
## [16] 29903 ATTAAAGGTTTATACCTTCCCAG...AAAAAAAAAAAAAAAAAAAAAA MT876433  
## [17] 29835 AACAAACCAACCAACTTTCGATC...ATGTGATTTTAATAGCTTCTTC MT845878  
## [18] 29815 ACTTTCGATCTCTTGTAGATCTG...TATCCCCATGTGATTTTAATAG MT919790  
## [19] 29852 ATTAAAGGTTTATACCTTCCCAG...CTATCCCCATGTGATTTTAATA MT955161  
## [20] 29825 ATACCTTCCCAGGTAACAAACCA...ATTAATTTTAGTAGTGCTATCC MT820489

Orienta los nucleótidos de los genomas con la función OrientNucleotides del paquete DECIPHER y muestra el código empleado para hacerlo.

oriented\_dna\_string\_set <- OrientNucleotides(my\_dna\_string\_set)

## ========================================================================================================================================================================================================  
##   
## Time difference of 0.18 secs

oriented\_dna\_string\_set

## DNAStringSet object of length 20:  
## width seq names   
## [1] 29866 AACTTTCGATCTCTTGTAGATCT...AAAAAAAAAAAAAAAAAAAAAA MW133981  
## [2] 29903 ATTAAAGGTTTATACCTTCCCAG...AAAAAAAAAAAAAAAAAAAAAA MT577009  
## [3] 29858 ATGTTTATACCTTCCCAGGTAAC...TGATTTTAATAGCTTCTTAGGA MT835383  
## [4] 29903 ATTAAAGGTTTATACCTTCCCAG...AAAAAAAAAAAAAAAAAAAAAA MT890462  
## [5] 29826 AACAAACCAACCAACTTTCGATC...GCTATCCCCATGTGATTTTAAT MW056032  
## ... ... ...  
## [16] 29903 ATTAAAGGTTTATACCTTCCCAG...AAAAAAAAAAAAAAAAAAAAAA MT876433  
## [17] 29835 AACAAACCAACCAACTTTCGATC...ATGTGATTTTAATAGCTTCTTC MT845878  
## [18] 29815 ACTTTCGATCTCTTGTAGATCTG...TATCCCCATGTGATTTTAATAG MT919790  
## [19] 29852 ATTAAAGGTTTATACCTTCCCAG...CTATCCCCATGTGATTTTAATA MT955161  
## [20] 29825 ATACCTTCCCAGGTAACAAACCA...ATTAATTTTAGTAGTGCTATCC MT820489

Realiza el alineamiento de las secuencias de los genomas virales con la función AlignSeqs y visualiza el resultado del alineamiento en tu navegador con la función BrowseSeqs de DECIPHER.

aligned\_dna\_string\_set <- AlignSeqs(oriented\_dna\_string\_set)

## Determining distance matrix based on shared 11-mers:  
## ================================================================================  
##   
## Time difference of 0.03 secs  
##   
## Clustering into groups by similarity:  
## ================================================================================  
##   
## Time difference of 0 secs  
##   
## Aligning Sequences:  
## ================================================================================  
##   
## Time difference of 0.41 secs  
##   
## Iteration 1 of 2:  
##   
## Determining distance matrix based on alignment:  
## ================================================================================  
##   
## Time difference of 0.01 secs  
##   
## Reclustering into groups by similarity:  
## ================================================================================  
##   
## Time difference of 0 secs  
##   
## Realigning Sequences:  
## ================================================================================  
##   
## Time difference of 0.44 secs  
##   
## Iteration 2 of 2:  
##   
## Determining distance matrix based on alignment:  
## ================================================================================  
##   
## Time difference of 0.01 secs  
##   
## Reclustering into groups by similarity:  
## ================================================================================  
##   
## Time difference of 0 secs  
##   
## Realigning Sequences:  
## ================================================================================  
##   
## Time difference of 0.01 secs

BrowseSeqs(aligned\_dna\_string\_set)

Guardar el resultado del alineamiento en un archivo en formato .fasta con la función writeXStringSet de Biostrings

writeXStringSet(aligned\_dna\_string\_set, "coronavirus\_sequences\_aligned.fasta", format="fasta")

Carga el archivo .fasta que construiste en el paso anterior con la función read.alignment de seqinr.

alignment <- read.alignment(file = "coronavirus\_sequences\_aligned.fasta", format = "fasta")  
  
head(alignment)

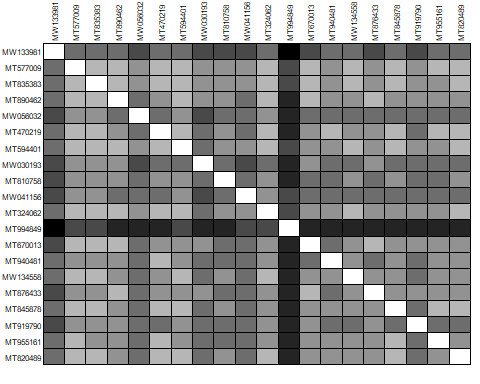
## $nb  
## [1] 20  
##   
## $nam  
## [1] "MW133981" "MT577009" "MT835383" "MT890462" "MW056032" "MT470219"  
## [7] "MT594401" "MW030193" "MT810758" "MW041156" "MT324062" "MT994849"  
## [13] "MT670013" "MT940481" "MW134558" "MT876433" "MT845878" "MT919790"  
## [19] "MT955161" "MT820489"  
##   
## $seq  
## $seq[[1]]  
## [1] "------------------------------------- -------------------------------------------------------------------"  
##   
##   
## $com  
## [1] NA

Crear una matriz de distancia con la función dist.alignment de seqinr y obtén una tabla en escala de grises con la función table.paint del paquete ape4 en donde sombras más oscuras de gris significan una mayor distancia. Muestra el contenido de la matriz de distancia que creaste, la imagen de la tabla en escala de grises que construiste, y el código empleado para conseguir todo lo anterior.

dist\_matrix <- dist.alignment(alignment, matrix = "similarity")  
  
dist\_matrix

## MW133981 MT577009 MT835383 MT890462 MW056032 MT470219  
## MT577009 0.02245783   
## MT835383 0.02247030 0.01002375   
## MT890462 0.02460132 0.01293087 0.01417572   
## MW056032 0.02782824 0.01831060 0.01831060 0.02087729   
## MT470219 0.02090710 0.01156572 0.01294061 0.01530001 0.02005825   
## MT594401 0.02321426 0.01003283 0.01003283 0.01418856 0.01921140 0.01295232  
## MW030193 0.02796700 0.01930359 0.01930359 0.01646216 0.02536989 0.02098522  
## MT810758 0.02612873 0.01649320 0.01649320 0.01933999 0.02332490 0.01843996  
## MW041156 0.02658722 0.01829711 0.01919402 0.02086191 0.02456624 0.02004347  
## MT324062 0.02319435 0.01001621 0.01157443 0.01416505 0.01920432 0.01293087  
## MT994849 0.03576896 0.02895499 0.02895499 0.03064308 0.03327345 0.03009091  
## MT670013 0.02530653 0.01419380 0.01419380 0.01003653 0.02168139 0.01638959  
## MT940481 0.02392187 0.01637887 0.01637887 0.01920593 0.02316633 0.01418452  
## MW134558 0.02462452 0.01295189 0.01295189 0.01638299 0.02088430 0.01532488  
## MT876433 0.02593207 0.01530001 0.01636872 0.01156572 0.02242581 0.01734858  
## MT845878 0.02170837 0.01418119 0.01418119 0.01736834 0.02087729 0.01157889  
## MT919790 0.02782777 0.01831397 0.01831397 0.01532257 0.02457160 0.02006195  
## MT955161 0.02247295 0.01417715 0.01531435 0.01736339 0.02166539 0.01157559  
## MT820489 0.02321622 0.01294776 0.01294776 0.01637777 0.02088254 0.01294776  
## MT594401 MW030193 MT810758 MW041156 MT324062 MT994849  
## MT577009   
## MT835383   
## MT890462   
## MW056032   
## MT470219   
## MT594401   
## MW030193 0.02016195   
## MT810758 0.01749368 0.02404276   
## MW041156 0.01737737 0.02536989 0.02258424   
## MT324062 0.01158491 0.02016195 0.01749368 0.01919017   
## MT994849 0.02838045 0.03343480 0.03140211 0.02895499 0.02952842   
## MT670013 0.01533106 0.01746076 0.02019996 0.02168139 0.01533106 0.03120485  
## MT940481 0.01737766 0.02399752 0.02181845 0.02316322 0.01737241 0.03224508  
## MW134558 0.01418975 0.02177739 0.01933999 0.02088430 0.01418808 0.03065697  
## MT876433 0.01638354 0.01840525 0.02102478 0.02240929 0.01635640 0.03171861  
## MT845878 0.01418856 0.02177739 0.01933999 0.02166212 0.01531743 0.03118705  
## MT919790 0.01921172 0.02098522 0.02332490 0.02457077 0.01920786 0.03327233  
## MT955161 0.01532540 0.02254174 0.02019996 0.02165595 0.01531307 0.03119124  
## MT820489 0.01418975 0.02177739 0.01933999 0.02087764 0.01418356 0.03065697  
## MT670013 MT940481 MW134558 MT876433 MT845878 MT919790  
## MT577009   
## MT835383   
## MT890462   
## MW056032   
## MT470219   
## MT594401   
## MW030193   
## MT810758   
## MW041156   
## MT324062   
## MT994849   
## MT670013   
## MT940481 0.02007307   
## MW134558 0.01738378 0.01921334   
## MT876433 0.01295711 0.02087904 0.01831674   
## MT845878 0.01738378 0.01532103 0.01638299 0.01920142   
## MT919790 0.01638959 0.02316555 0.02088710 0.01737416 0.02088114   
## MT955161 0.01832412 0.01638079 0.01737678 0.01919595 0.01294733 0.02166975  
## MT820489 0.01738378 0.01737912 0.01638299 0.01831090 0.01418689 0.02088710  
## MT955161  
## MT577009   
## MT835383   
## MT890462   
## MW056032   
## MT470219   
## MT594401   
## MW030193   
## MT810758   
## MW041156   
## MT324062   
## MT994849   
## MT670013   
## MT940481   
## MW134558   
## MT876433   
## MT845878   
## MT919790   
## MT955161   
## MT820489 0.01532000

temp <- as.data.frame(as.matrix(dist\_matrix))  
  
  
grayscale\_table <- table.paint(temp, cleg = 0, clabel.row = .5, clabel.col = .5) + scale\_color\_viridis()

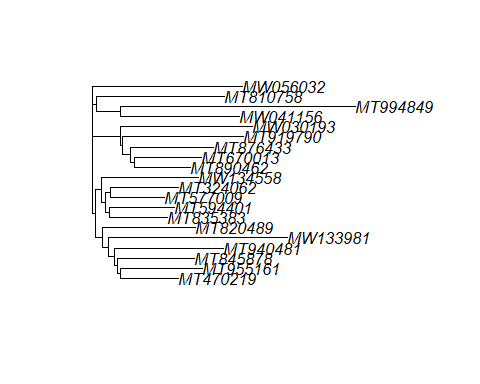


Construye un objeto de tipo phylo con la función nj del paquete ape a partir de la matriz de distancia que obtuviste en el paso anterior. Muestra el código que empleaste para hacerlo

virus\_tree <- nj(dist\_matrix)  
class(virus\_tree)

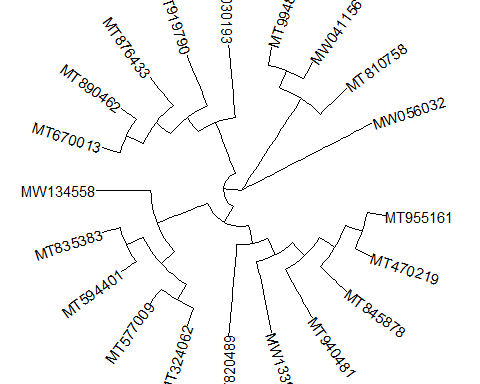
## [1] "phylo"

virus\_tree <- ladderize(virus\_tree)  
plot(virus\_tree)



Finalmente, construye un árbol filogenético utilizando el objeto de tipo phylo que construiste en el paso anterior y la función plot del código base de R. Muestra la imagen de tu árbol filogenético y el código que empleaste para obtenerlo.

ggtree(virus\_tree, branch.length = "none", layout = "circular") + geom\_tiplab()



plot\_virus\_filogenia <- ggtree(virus\_tree) + geom\_tiplab() + ggtitle("Analisis filogenetico de los genomas del SARS-CoV")  
  
plot\_virus\_filogenia

