

Phylogenetic Analysis of Cordyceps using ITS sequences

Paul Arturo Gamez Coronel

2023-10-07

Introducción

En este documento, documentaré el análisis de secuencias de Cordyceps utilizando R. El análisis incluye la instalación de paquetes, el procesamiento de secuencias y la construcción de un árbol filogenético.

Configuración y preparación

```
setwd("C:/Users/Paulw/Desktop/Arbol filogenetico/cordyceps data")
```

```
library(seqinr)
```

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

```
library(adegenet)
```

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

```
## Loading required package: ade4
```

```
## Warning: package 'ade4' 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
```

```
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
##
## S Xu, Z Dai, P Guo, X Fu, S Liu, L Zhou, W Tang, T Feng, M Chen, L
## Zhan, T Wu, E Hu, Y Jiang, X Bo, G Yu. ggtreeExtra: Compact
## visualization of richly annotated phylogenetic data. Molecular Biology
## and Evolution. 2021, 38(9):4039-4042. doi: 10.1093/molbev/msab166
##
## 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
```

```
library(DECIPHER)
```

```
## Loading required package: Biostrings
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following object is masked from 'package:ade4':
##
##   score
```

```
## 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 object is masked from 'package:ggtree':
##
##   expand

## 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:ggtree':
##
##   collapse

## 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:ape':
##
##   complement

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

```

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

## Loading required package: RSQLite

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

## Loading required package: parallel
```

```
library(viridis)
```

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

## Loading required package: viridisLite

## Warning: package 'viridisLite' was built under R version 4.2.3
```

```
library(ggplot2)
```

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

Inicializar los primeros pasos

en esta parte del código estableceremos un archivo para trabajar con la variable “seq_name” donde a partir de el y usando funciones de la librería ape leeremos la secuencia de formato fasta y realizaremos un alineamiento de las mismas.

```
seq_name <- file.choose()
seqs <- readDNAStringSet("cordyceps.txt", format = "fasta")
seqs <- OrientNucleotides(seqs)
```

```
## =====
##
## Time difference of 0.08 secs
```

```
aligned <- AlignSeqs(seqs)
```

```
## Determining distance matrix based on shared 9-mers:
## =====
##
## Time difference of 0 secs
##
## Clustering into groups by similarity:
## =====
##
## Time difference of 0.01 secs
##
```

```

## Aligning Sequences:
## =====
##
## Time difference of 0.23 secs
##
## Iteration 1 of 2:
##
## Determining distance matrix based on alignment:
## =====
##
## Time difference of 0 secs
##
## Reclustering into groups by similarity:
## =====
##
## Time difference of 0.01 secs
##
## Realigning Sequences:
## =====
##
## Time difference of 0.2 secs
##
## Iteration 2 of 2:
##
## Determining distance matrix based on alignment:
## =====
##
## Time difference of 0 secs
##
## Reclustering into groups by similarity:
## =====
##
## Time difference of 0.01 secs
##
## Realigning Sequences:
## =====
##
## Time difference of 0.03 secs

```

Primera visualizacion

Con esta funcion podremos observar en nuestro explorador una visualizacion basica de las secuencias donde podremos comparar las secuencias de ADN.

```
BrowseSeqs(aligned, colorPatterns = TRUE, highlight = 0)
```

Guardar secuencia alineada

Guardaremos la secuencia con un nombre a eleccion propia

```
writeXStringSet(aligned, file = "Cordyceps.fasta")
```

Analisis de distancia

esta seccion del codigo nos permitira leer a partir del archivo creado anteriormente, el distanciamiento en funcion al alineamiento, que compara la similitud en las secuencias

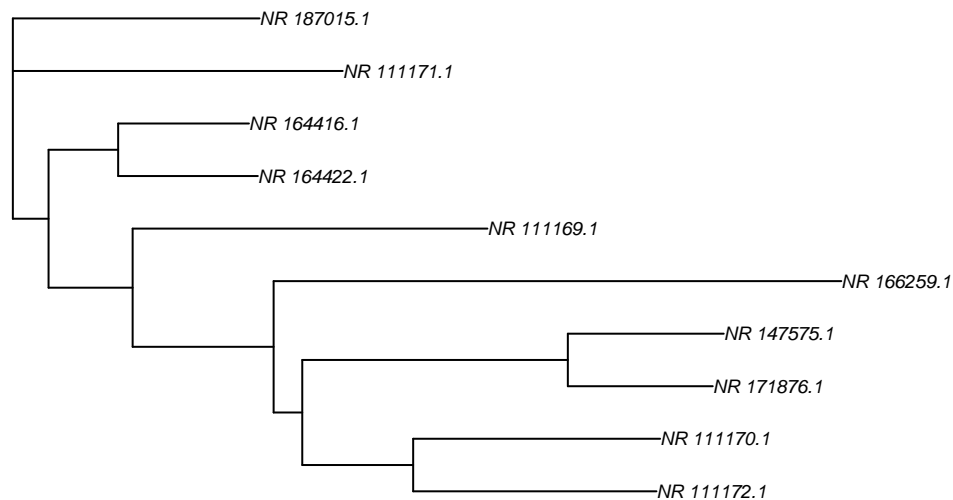
```
dna <- read.alignment("Cordyceps.fasta", format = "fasta")
D <- dist.alignment(dna, matrix = "similarity")
tre <- nj(D)
tre <- ladderize(tre)
```

Graficas de la similitud

Aqui podemos observar una grafica de la matriz creada en el codigo anterior, cada region nos indica en funcion del color la similitud, mientras mas oscuro sea una casilla representa un mayor distanciamiento entre las secuencias comparadas, caso contrario al color blanco que nos dice que tiene un menor distanciamiento

```
plot(tre, cex = 0.6)
title("Similitud en Cordyceps de region (ITS)")
```

Similitud en Cordyceps de region (ITS)



Dendrograma

los dendogramas nos permiten visualizar de una manera mas clara tanto el distanciamiento de las secuencias como su similitud

```
h_cluster <- hclust(D, method = "complete", members = NULL)
plot(h_cluster, cex = 0.6, main = "Dendrograma", xlab = "Distancia", ylab = "Similitud")
```

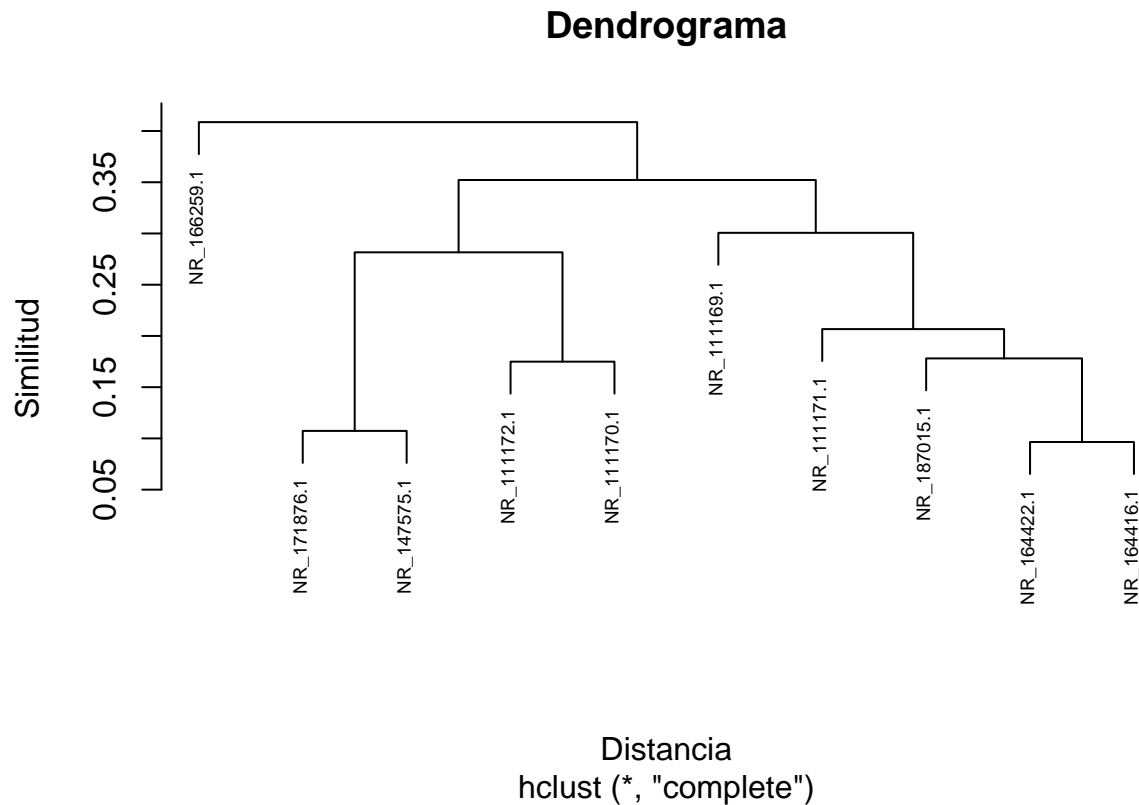


Grafico usando GGTREE

este codigo nos permite usar la libreria ggtree para generar un arbol filogenetico, a diferencia de las graficas con R base esta nos permite visualizar de una manera mas personalizada nuestro arbol.

```
ggtree(tre) +
  geom_tiplab(hjust = -0.3, size = 4, align = TRUE) +
  xlim(0, 0.5) +
  ggtitle("Árbol filogenético de Cordyceps de regiones ITS")
```

Árbol filogenético de Cordyceps de regiones ITS

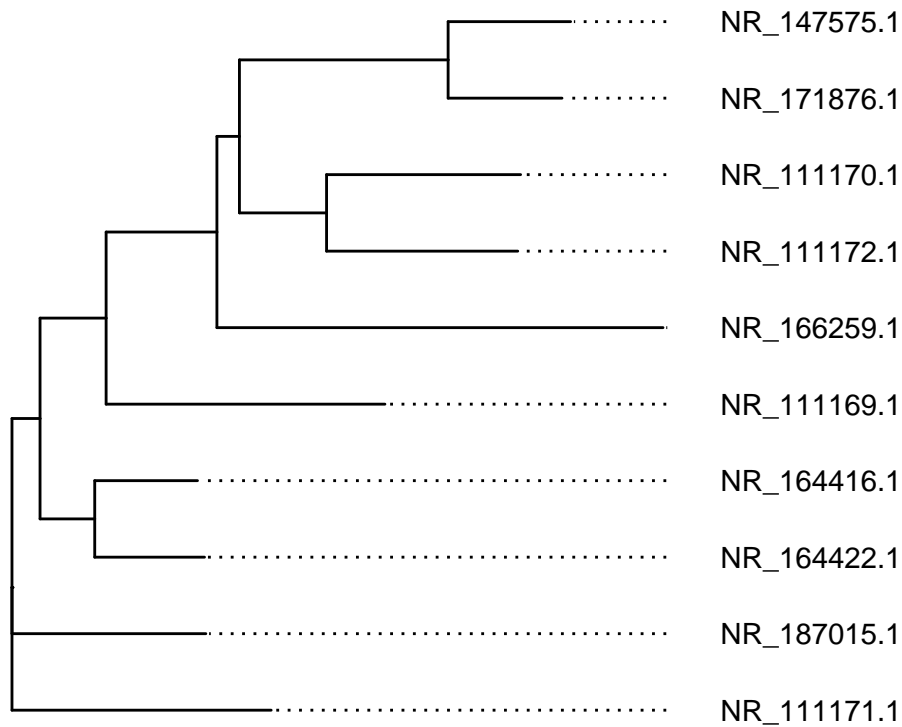


Grafico del alineamiento

Aquí podemos observar un grafico del alineamiento de la secuencia fasta donde podemos observar las bases nitrogenadas de las secuencias.

```
tre.new <- tre
tre.new$tip.label <- aligned@ranges@NAMES

msaplot(p = ggtree(tre.new), fasta = "Cordyceps.fasta", window = c(150, 175)) +
  scale_fill_viridis_d(alpha = 0.8)
```

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```