Creating phylogenetic trees

Paul Cool

2024-05-05

```
library(reticulate)
# directory with files
directory = '/Users/paulcool/Dropbox/cool_functions/phylogenetic_tree/'
```

Phylogenetic trees

A phylogenetic tree is a diagram that shows the descent of different species, organisms, or genes from a common ancestor.

Finding Taxonomy ID

https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=1280&lvl=3&lin=f&keep=1&srchmode=1&unlock

```
# define bacteriae
staphylococcus_aureus = 1280
staphylococcus_epidermidis = 1282
staphylococcus_haemolyticus = 1283
staphylococcus_hominis = 1290
staphylococcus_warneri = 1292
streptococcus = 1301
enterococcus = 1350
cutibacterium_acnes = 1747
staphylococcus_lugdunensis = 28035
escherichia_coli = 562
shigella = 620
# save as vector
pji_bact = c(staphylococcus_aureus, staphylococcus_epidermidis,
```

```
            staphylococcus_haemolyticus, staphylococcus_hominis,
            staphylococcus_warneri, streptococcus, enterococcus,
            cutibacterium_acnes, staphylococcus_lugdunensis,
            escherichia_coli, shigella)
print(pji_bact)
```

```
 [1]  1280  1282  1283  1290  1292  1301  1350  1747 28035   562   620
```

```
# convert to data frame, including directory name for reading python
df = tibble(pji_bact, directory)
# write to csv
write.csv(df, file=paste(directory,'pji_bact.csv', sep=""),
          row.names=FALSE)
```

A csv file with the pji bacteriae is now created in the defined directory.

Downloading NCBI taxonomy database

The NCBI taxonomy database can be downloaded from:

https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/

Subsequently, unzip the file. It will create a directory with several dmp files. These files can be read into R to show classification.

```
      1 |...2                       root |...4 ...5 |...6          ...7
1    562     |           Escherichia coli     | <NA>     |    Escherichia
2    620     |                   Shigella     | <NA>     |           <NA>
3   1747     |        Cutibacterium acnes     | <NA>     |  Cutibacterium
4   1280     |      Staphylococcus aureus     | <NA>     | Staphylococcus
5   1282     |  Staphylococcus epidermidis    | <NA>     | Staphylococcus
6   1283     | Staphylococcus haemolyticus    | <NA>     | Staphylococcus
7   1290     |     Staphylococcus hominis     | <NA>     | Staphylococcus
8  28035     |  Staphylococcus lugdunensis    | <NA>     | Staphylococcus
9   1292     |     Staphylococcus warneri     | <NA>     | Staphylococcus
10  1350     |               Enterococcus     | <NA>     |           <NA>
11  1301     |              Streptococcus     | <NA>     |           <NA>
   |...8                ...9 |...10              ...11 |...12
1      |   Enterobacteriaceae      |    Enterobacterales      |
2      |   Enterobacteriaceae      |    Enterobacterales      |
3      | Propionibacteriaceae      | Propionibacteriales      |
```

```
4      |    Staphylococcaceae      |          Bacillales      |
5      |    Staphylococcaceae      |          Bacillales      |
6      |    Staphylococcaceae      |          Bacillales      |
7      |    Staphylococcaceae      |          Bacillales      |
8      |    Staphylococcaceae      |          Bacillales      |
9      |    Staphylococcaceae      |          Bacillales      |
10     |      Enterococcaceae      |     Lactobacillales      |
11     |     Streptococcaceae      |     Lactobacillales      |
                  ...13 |...14          ...15 |...16 ...17 |...18    ...19
1  Gammaproteobacteria      | Pseudomonadota      |  <NA>      | Bacteria
2  Gammaproteobacteria      | Pseudomonadota      |  <NA>      | Bacteria
3        Actinomycetes      | Actinomycetota      |  <NA>      | Bacteria
4              Bacilli      |      Bacillota      |  <NA>      | Bacteria
5              Bacilli      |      Bacillota      |  <NA>      | Bacteria
6              Bacilli      |      Bacillota      |  <NA>      | Bacteria
7              Bacilli      |      Bacillota      |  <NA>      | Bacteria
8              Bacilli      |      Bacillota      |  <NA>      | Bacteria
9              Bacilli      |      Bacillota      |  <NA>      | Bacteria
10             Bacilli      |      Bacillota      |  <NA>      | Bacteria
11             Bacilli      |      Bacillota      |  <NA>      | Bacteria
   |...20
1       |
2       |
3       |
4       |
5       |
6       |
7       |
8       |
9       |
10      |
11      |
```

Python Tree from NCBI database

Using the taxonomy identification numbers, the Python ete3 package can download data from the NCBI site and save this as a Newick file. The tree can than be plotted in R using the Bioconductor **ggtree** package (https://bioconductor.org/packages/release/bioc/html/ggtree.html) (based on ggplot2).

```
from pandas import read_csv
from ete3 import NCBITaxa
```

```
ncbi = NCBITaxa()

data = read_csv('pji_bact.csv')
bact = data['pji_bact'].tolist()
directory = data['directory'][0]
# print(bact)
# print(directory)

tree = ncbi.get_topology(bact)

for node in tree.traverse():
    node.name = node.sci_name # insert scientific names

print(tree.write())
```

```
((((Enterococcus:1,Streptococcus:1)1:1,(Staphylococcus aureus:1,Staphylococcus epidermidis:1
```

```
# save tree to file
tree.write(features=["node.name"], outfile=f'{directory}pji_bact_tree.nwk')
```

The Newick file should now be created in the directory specified.

BiocManager

To Install Bioconductor and the ggtree package:

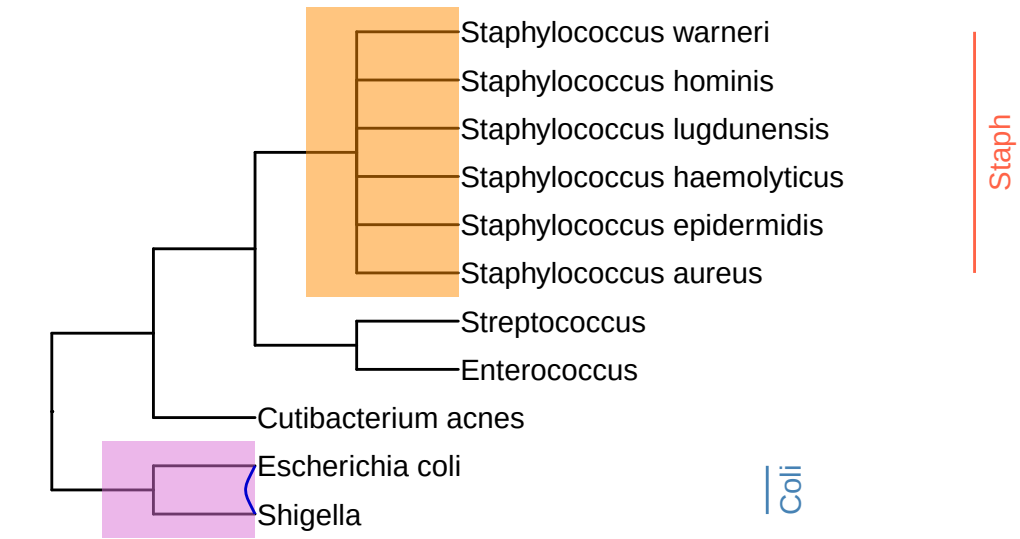
```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("ggtree")
```

For information re the ggtree package:

https://4va.github.io/biodatasci/r-ggtree.html

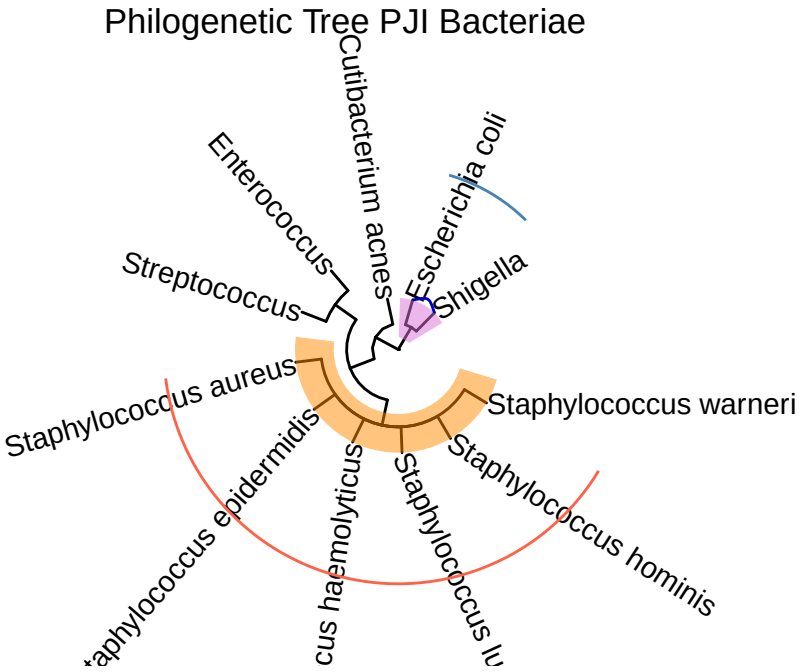
https://yulab-smu.top/treedata-book/chapter4.html

```
library(ggtree)

# read with phytools as read.table removes whitespace in names
tree = phytools::read.newick(paste(directory,'pji_bact_tree.nwk',
                                   sep=''))

#tree
# show the tree
ggtree(tree) +
  geom_tiplab() +
  theme_tree() +
  #geom_treescale(fontsize=3, linesize=1, offset=4) + #
  geom_cladelabel(node=16, label='Staph',
                  color='tomato', offset=5, angle=90, vjust=1.5, hjust=0.5) +
  geom_cladelabel(node=17, label='Coli',
                  color='steelblue', offset=5, angle=90, vjust=1.5, hjust=0.5) +
  geom_hilight(node=16, fill='darkorange') +
  geom_hilight(node=17, fill='orchid') +
  # insert a taxonomic link
  geom_taxalink(taxa1='Escherichia coli', taxa2='Shigella', color='blue3') +
  ggtitle('Philogenetic Tree PJI Bacteriae')
```



Or, a more fancy tree (see documentation for more options).

```
ggtree(tree, layout='circular') +
  geom_tiplab() +
  theme_tree() +
  #geom_treescale(fontsize=3, linesize=1, offset=4) + #
  geom_cladelabel(node=16, label='',
                  color='tomato', offset=5, angle=90, vjust=1.5, hjust=0.5) +
  geom_cladelabel(node=17, label='',
                  color='steelblue', offset=5, angle=90, vjust=1.5, hjust=0.5) +
  geom_hilight(node=16, fill='darkorange') +
  geom_hilight(node=17, fill='orchid') +
  # insert a taxonomic link
  geom_taxalink(taxa1='Escherichia coli', taxa2='Shigella', color='blue3') +
  ggtitle('Philogenetic Tree PJI Bacteriae')
```



Make tree from FASTA READS

It is also possible to create phylogenetic trees from fasta files. The sequences (DNA or protein) are compared for similarity using the msa (multiple sequence alignment) package (https://bioconductor.org/packages/release/bioc/html/msa.html, https://bioconductor.org/packages/devel/bioc/vignettes/msa/inst/doc/msa.pdf).

To install:

```r
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("seqinr")
BiocManager::install("ape")
BiocManager::install("msa")
```

```r
library(seqinr)
library(ape)
library(msa)

## Read sequences from FASTA file of the cases
sequence = readAAStringSet(paste(directory,'input_1.fasta',
                                 sep=''))
sequence # see sequence
```

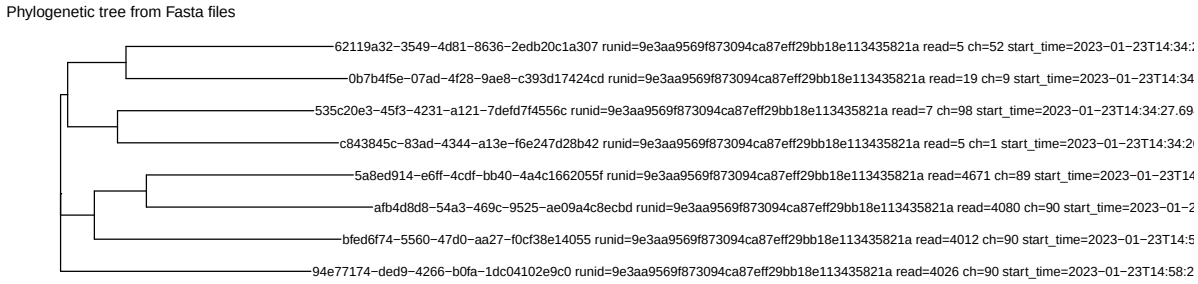
```
AAStringSet object of length 8:
    width seq                                               names
[1]   338 TTGTACTTCGTTCAATTTCATTT...ACTGCGCTGACCGGATCCAGCTG 62119a32-3549-4d8...
[2]   252 AAACACTTCGTTCAGTTGCCCCG...AATTGCTGATTATAGGAGTAGAG 535c20e3-45f3-423...
[3]   518 AGTACTTCGTTCAGCTCAGATGC...CTGAGCTCCAGAAAGTAGGTTCT c843845c-83ad-434...
[4]   915 TTATTACTTCGTTCAGTTACAGA...TGATGTAGGCATTTTAATGCTAT 0b7b4f5e-07ad-4f2...
[5]   303 GTTAGCCGTTCAGTTGGAGCTGG...TGGGTATATACCCAGTAATGGGA 5a8ed914-e6ff-4cd...
[6]   305 AATACTTCGTTCAGTTTCGGAAG...CAAGGCAGTAAATATATTGACTT bfed6f74-5560-47d...
[7]   286 GTATGCTTCGTTCATTGGAAGTG...TAAATGTGGAATTAATAATTAAT 94e77174-ded9-426...
[8]   801 GGTGTACTTCGTTTATTTCAGAT...AAAGTCTTTTTAAAAGAATGCGA afb4d8d8-54a3-469...
```

```r
## Perform multiple sequence alignment
alignment = msa(sequence) # this takes a while if there are many reads, best to limit to a fe
```

```
use default substitution matrix
```

```r
## Compute distance matrix
alignment_sequence = msaConvert(alignment, type="seqinr::alignment")
distance_alignment = dist.alignment(alignment_sequence)

## phylogenetic tree using neighbour joining
tree = bionj(distance_alignment)

## display phylogenetic tree
```

```
ggtree(tree) +
  geom_tiplab() +
  theme_tree() +
  geom_cladelabel(node=6, label='',
                  color='tomato', offset=1, angle=90, vjust=1.5,
                  hjust=0.5) +
  ggtitle('Phylogenetic tree from Fasta files')
```



The tree can be further improved as required using the ggtree package.

Packages

Table 1: Packages Used

package	version	date
base	4.4.0	2024-04-24
knitr	1.46	2024-04-05
msa	1.36.0	2024-04-26
Biostrings	2.72.0	2024-05-01
GenomeInfoDb	1.40.0	2024-05-01
XVector	0.44.0	2024-05-01
IRanges	2.38.0	2024-05-01
S4Vectors	0.42.0	2024-05-01
BiocGenerics	0.50.0	2024-05-01
ape	5.8	2024-04-09
seqinr	4.2.36	2023-12-08
ggtree	3.12.0	2024-05-01
reticulate	1.36.1	2024-04-22
survival	3.6.4	2024-04-22
colorspace	2.1.0	2023-01-23
readxl	1.4.3	2023-07-05
lubridate	1.9.3	2023-09-24

package	version	date
forcats	1.0.0	2023-01-27
stringr	1.5.1	2023-11-14
dplyr	1.1.4	2023-11-16
purrr	1.0.2	2023-08-08
readr	2.1.5	2024-01-10
tidyr	1.3.1	2024-01-23
tibble	3.2.1	2023-03-19
ggplot2	3.5.1	2024-04-22
tidyverse	2.0.0	2023-02-21