Calculating and plotting pairwise distances in DNA alignments

Silvia Justi

10/3/2017

Setting work directory

```
setwd("/Users/silviajusti/Desktop")
```

Installing and Loading the necessary libraries

```
#install.packages()
library(ape)
```

Reading the fasta files. For this one I am using 5 random Hemitera 18S sequences. Sample names are GenBank access numbers

```
data <-read.FASTA("sequence.fasta")
data
```

```
## 5 DNA sequences in binary format stored in a list.
##
## All sequences of same length: 2538
##
## Labels:
## KY554599
## U06481
## U06482
## U06480
## U06479
##
## Base composition:
##     a     c     g     t
## 0.219 0.259 0.294 0.228
```

```
summary(data)
```

```
##          Length Class  Mode
## KY554599 2538   -none- raw
## U06481   2538   -none- raw
## U06482   2538   -none- raw
## U06480   2538   -none- raw
## U06479   2538   -none- raw
```

Calculating the pairwise k2p distances between samples and making sure they are returned as matrix, so we can plot the distances later and saving as tables

```
dist <-dist.dna(data, model ="K80", as.matrix = TRUE, pairwise.deletion = T)
summary(dist)
```

```
##     KY554599              U06481              U06482              U06480
##  Min.   :0.0000    Min.   :0.00000    Min.   :0.0000    Min.   :0.00000
##  1st Qu.:0.1250    1st Qu.:0.09182    1st Qu.:0.1616    1st Qu.:0.09182
##  Median :0.1335    Median :0.12500    Median :0.1652    Median :0.13350
##  Mean   :0.1414    Mean   :0.13952    Mean   :0.1681    Mean   :0.13025
##  3rd Qu.:0.1652    3rd Qu.:0.18432    3rd Qu.:0.1843    3rd Qu.:0.16161
##  Max.   :0.2831    Max.   :0.29645    Max.   :0.3296    Max.   :0.26432
##      U06479
##  Min.   :0.0000
##  1st Qu.:0.2643
##  Median :0.2831
##  Mean   :0.2347
##  3rd Qu.:0.2964
##  Max.   :0.3296
```

I am now ramdomly choosing a sample to compare to all others

```
sample <- subset(dist, select = c("KY554599"))
```

Ploting the distances between the chosen sample and the others

```
sample_sorted <- as.matrix(sample[order(sample[,1]),])
plot (sample_sorted, col = rgb(1,0.4,0), pch=18, main = "K2P distances", xaxt="n", xlab
= "" , ylab="distance from KY554599")
axis(1, at=1:5, labels = row.names(sample_sorted), las =2, cex.axis = 0.75)
```

K2P distances