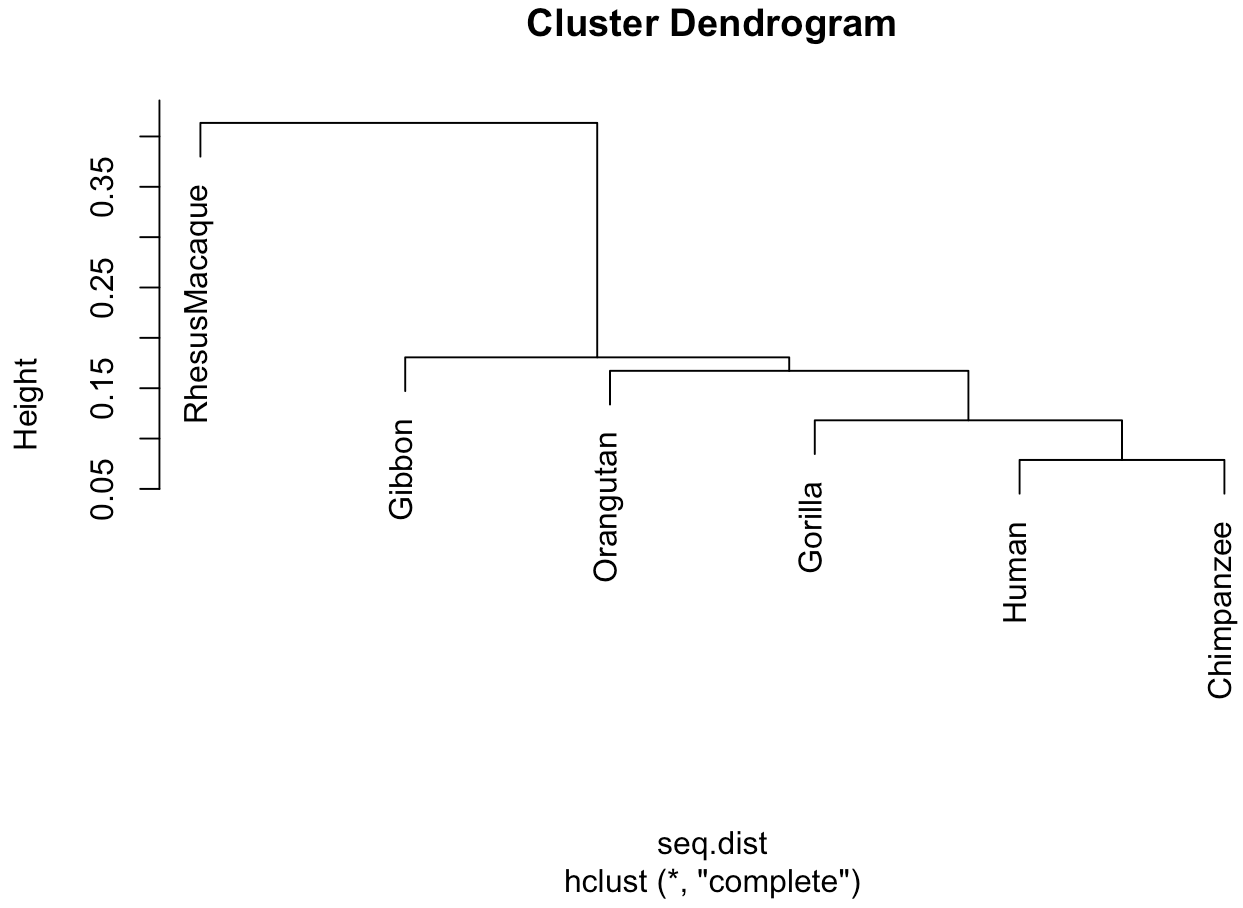
Lab5_2

2024-05-02

```
library(seqinr)
#4
seq.align <- read.alignment("PrimateAlignN.txt", format = "fasta")
#5
seq.dist <- dist.alignment(seq.align)
seq.dist
```

```
##                  Gibbon  Orangutan    Gorilla      Human Chimpanzee
## Orangutan     0.18064892
## Gorilla       0.17827772 0.16724840
## Human         0.16935811 0.15267620 0.11812488
## Chimpanzee    0.16935811 0.15267620 0.11812488 0.07874992
## RhesusMacaque 0.40919660 0.41299001 0.41344912 0.40872383 0.40681834
```

```
#6
seq.clust <- hclust(seq.dist)
#7
plot(seq.clust)
```



```
#8
cutree(tree=seq.clust, k=3)
```

```
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             2             2             2
## RhesusMacaque
##             3
```

```
#9,10
kmeans(x=seq.dist,centers=3)
```

```
## K-means clustering with 3 clusters of sizes 2, 3, 1
##
## Cluster means:
##       Gibbon  Orangutan    Gorilla      Human Chimpanzee RhesusMacaque
## 1 0.09032446 0.09032446 0.17276306 0.16101716 0.16101716     0.4110933
## 2 0.17233132 0.15753360 0.07874992 0.06562494 0.06562494     0.4096638
## 3 0.40919660 0.41299001 0.41344912 0.40872383 0.40681834     0.0000000
##
## Clustering vector:
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             2             2             2
## RhesusMacaque
##             3
##
## Within cluster sum of squares by cluster:
## [1] 0.03298034 0.02399052 0.00000000
##  (between_SS / total_SS =  90.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
#11
kmeans(x=seq.dist, centers=3)$cluster
```

```
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             2             2             3             3             3
## RhesusMacaque
##             1
```

```
# 13
for(j in 1:3){
  print(j)
}
```

```
## [1] 1
## [1] 2
## [1] 3
```

```
#14, 15
for(Nclusters in 1:5){
  print( cutree(tree=seq.clust, k=Nclusters) )
  print( kmeans(x=seq.dist, centers=Nclusters)$cluster )
}
```

```
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             1
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             1
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             2
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             2
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             2             2             2
## RhesusMacaque
##             3
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             3             3             2             2             2
## RhesusMacaque
##             1
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             3             3             3
## RhesusMacaque
##             4
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             3             2             1             1             1
## RhesusMacaque
##             4
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             3             4             4
## RhesusMacaque
##             5
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             4             1             5             3             3
## RhesusMacaque
##             2
```

```
#16
for (Nclusters in 1:5) {
  writeLines(paste("\n\nhclust # of clusters:", Nclusters))
  print( cutree(tree=seq.clust, k=Nclusters) )
  writeLines(paste("kmeans # of clusters:", Nclusters))
  print( kmeans(x=seq.dist, centers=Nclusters)$cluster )
}
```

```
##
##
## hclust # of clusters: 1
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             1
## kmeans # of clusters: 1
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             1
##
##
## hclust # of clusters: 2
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             2
## kmeans # of clusters: 2
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             2
##
##
## hclust # of clusters: 3
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             2             2             2
## RhesusMacaque
##             3
## kmeans # of clusters: 3
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             2             2             2
## RhesusMacaque
##             3
##
##
## hclust # of clusters: 4
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             3             3             3
## RhesusMacaque
##             4
## kmeans # of clusters: 4
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             2             3             1             1             1
## RhesusMacaque
##             4
##
##
## hclust # of clusters: 5
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
```

```
##             1             2             3             4             4
## RhesusMacaque
##             5
## kmeans # of clusters: 5
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             3             5             2             4             4
## RhesusMacaque
##             1
```

```
#17, 18
library(seqinr)
cluster.fxn <- function(align.file) {
  seq.align <- read.alignment(file = align.file, format = "fasta")
  seq.dist <- dist.alignment(seq.align)
  print(seq.dist)
  seq.clust <- hclust(seq.dist)
  plot(seq.clust)
  for (Nclusters in 1:5) {
    writeLines(paste("\n\nhclust # of clusters:", Nclusters))
    print( cutree(tree=seq.clust, k=Nclusters) )
    writeLines(paste("kmeans # of clusters:", Nclusters))
    print( kmeans(x=seq.dist, centers=Nclusters)$cluster )
  }
}
#19
cluster.fxn("PrimateAlignN.txt")
```

```
##                  Gibbon  Orangutan    Gorilla      Human Chimpanzee
## Orangutan     0.18064892
## Gorilla       0.17827772 0.16724840
## Human         0.16935811 0.15267620 0.11812488
## Chimpanzee    0.16935811 0.15267620 0.11812488 0.07874992
## RhesusMacaque 0.40919660 0.41299001 0.41344912 0.40872383 0.40681834
```

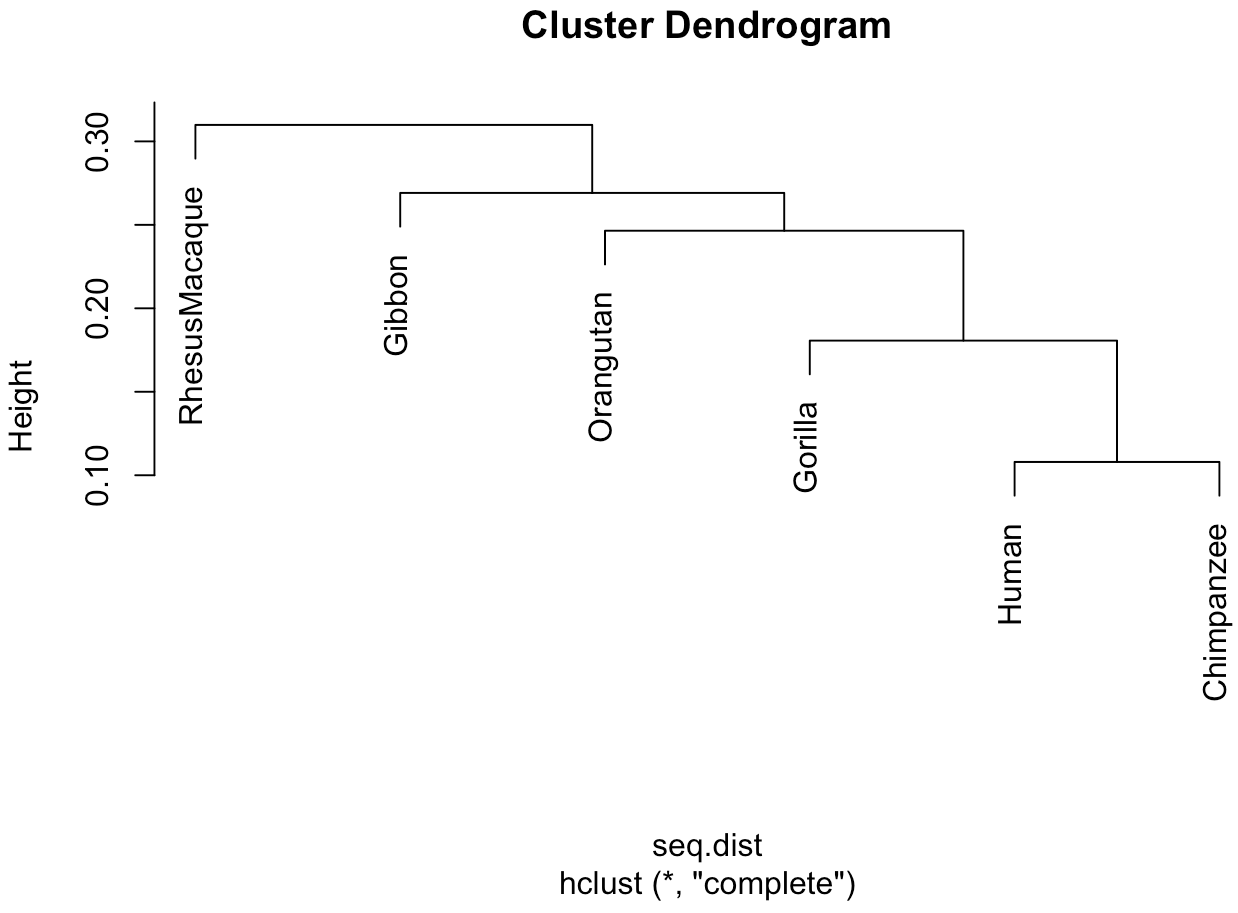

Cluster Dendrogram
Height
0.05
0.15
0.25
0.35
RhesusMacaque
Gibbon
Orangutan
Gorilla
Human
Chimpanzee
seq.dist
hclust (*, "complete")

```
##
##
## hclust # of clusters: 1
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             1
## kmeans # of clusters: 1
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             1
##
##
## hclust # of clusters: 2
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             2
## kmeans # of clusters: 2
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             2
##
##
## hclust # of clusters: 3
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             2             2             2
## RhesusMacaque
##             3
## kmeans # of clusters: 3
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             2             2             1             1             1
## RhesusMacaque
##             3
##
##
## hclust # of clusters: 4
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             3             3             3
## RhesusMacaque
##             4
## kmeans # of clusters: 4
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             3             2             4             4             4
## RhesusMacaque
##             1
##
##
## hclust # of clusters: 5
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
```

```
##             1             2             3             4             4
## RhesusMacaque
##             5
## kmeans # of clusters: 5
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             5             3             4             2             2
## RhesusMacaque
##             1
```

```
# 20
cluster.fxn("PrimateAlignAA.txt")
```

```
##                  Gibbon Orangutan   Gorilla     Human Chimpanzee
## Orangutan     0.2691280
## Gorilla       0.2508726 0.2464704
## Human         0.2365250 0.2215072 0.1740777
## Chimpanzee    0.2414023 0.2267198 0.1806489 0.1079584
## RhesusMacaque 0.2734344 0.3098689 0.2859645 0.2734344  0.2691280
```



```
##
##
## hclust # of clusters: 1
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             1
## kmeans # of clusters: 1
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             1
##
##
## hclust # of clusters: 2
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             1             1             1             1
## RhesusMacaque
##             2
## kmeans # of clusters: 2
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             2             2             2
## RhesusMacaque
##             1
##
##
## hclust # of clusters: 3
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             2             2             2
## RhesusMacaque
##             3
## kmeans # of clusters: 3
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             2             2             3             3             3
## RhesusMacaque
##             1
##
##
## hclust # of clusters: 4
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             1             2             3             3             3
## RhesusMacaque
##             4
## kmeans # of clusters: 4
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             4             3             1             1             1
## RhesusMacaque
##             2
##
##
## hclust # of clusters: 5
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
```

```
##             1             2             3             4             4
## RhesusMacaque
##             5
## kmeans # of clusters: 5
##        Gibbon     Orangutan       Gorilla         Human    Chimpanzee
##             2             5             4             3             3
## RhesusMacaque
##             1
```

Section 5: Assessment

Q1:

Human and Chimpanzee are most similar to each other.

RhesusMacaque and Gorilla are most different from one another as bigger values indciate more distance between sequences, which mean they are different from each other the most.

Q2:

Chimpanzee divered the lastest while RhesusMacaque diverged the earliest.

Q3:

It seems like the nucleotide sequence has higher dissimilarity compare to the amino acid sequences as one y axis reach to 0.35 while the amino acid sequence only reach to 0.3.

Also, for the nucleotide one, the difference between Gibbon, Orangutan, Gorilla, human and chimpanzee are reltaive small while the gap (distance in y axis) among these species for the amino acid are bigger. There is also a big difference between rhesusmacaque and all other species in the tree of nucleotide while the difference is smaller in the amino acid.

Q4:

As the gap is larger in nucleotide compare to the amino acid. There is redundancy in the genetic code, even though they have different codons, it could be synonymous substitution and cause to get a same amino acid. Thus, since amino acid only consider the nonsynonymous substitution in their branch length and nucleotide would consider both synonymous substitution and nonsynonymous substituion, the branch length is longer in nucleotide compare to the amino acid.

Q5:

It is better to use amino acid sequence when examining the protein function and structure as the nonsynomous subsitution will not influence the result in this case and we only want to focus on the overall amino acid sequence in this case.

It is better to use nucleotide sequence when we try to learn about the evolutionary relationship, neutral selection as we want to consider both nonsynomous substituion and synomous substitution for this type of problems.

Q6:

As a high Ka/Ks ratio suggests more non synonymous substitution, it could change the amino acid more. In this case, as there are more non synonymous substitution, the amino acid get changed and the sequences could get different species structure and thus different clustering amino acid.

Q7:

For nucleotide sequence, the value k = 3 is the value which two methods disagree.

For amino acid, the value k = 2 and k = 3 are the values we try to find.