

HW3a

Part 1 – Phylogenetic Trees

One can use the power of phylogenetic trees and alignment algorithms to visually assess distances between protein sequences while appreciating hierarchical structure.

In this HW we will calculate sequence alignment distances between different strains of Dengue virus strains. Specifically between *Dengue virus 1 NS1 protein (UniProt Q9YRR4)*, *Dengue virus 2 NS1 protein (UniProt Q9YP96)*, *Dengue virus 3 NS1 protein (UniProt B0LSS3)*, and *Dengue virus 4 NS1 protein (UniProt Q6TFL5)*. The main question is which of those strains are the mostly related? This information could be used to better understand dangers of unknown viral strains most similar to the known strain.

- 1) Retrieve sequences programmatically and save into one *.fasta file

```
library("seqinr");          # Load the SeqinR package
# Make a vector containing the names of the sequences
seqnames <- c("Q9YRR4", "Q9YP96", "B0LSS3", "Q6TFL5");
seqs <- retrieveseqs(seqnames, "swissprot");
```

- 2) Align sequences via the online tool [Clustal Omega](#), choose PHYLIP output. E.g. save file DengueViruses.phylip

The screenshot shows the Clustal Omega web interface. At the top, there's a teal header with the logo and navigation links: 'Input form', 'Web Services', and 'Help & Documentation'. Below the header, there's a breadcrumb trail: 'Tools > Multiple Sequence Alignment > Clustal Omega'. The main heading is 'Multiple Sequence Alignment', followed by a brief description: 'Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.' The interface is divided into three steps: Step 1 - Enter your input sequences, Step 2 - Set your parameters, and Step 3 - Submit your job. Step 1 includes a text area with protein sequences and a 'Choose File' button. Step 2 shows the 'OUTPUT FORMAT' set to 'PHYLIP'. Step 3 has a checkbox for 'Be notified by email'.

- 3) Calculate distances between sequences

```
NS1aln <- read.alignment(file = "Dengueviruses.phylip", format =
"phylip")
```

GBIO0009-1

```
NS1dist <- dist.alignment(NS1aln)
```

4) Build hierarchical tree. Which of the viruses is closest to Q6TFL5?

```
plot(hclust(NS1dist))
```

Appendix of auxiliary functions

```
retrieveseqs <- function(seqnames, acnucdb)
{
  myseqs <- list() # Make a list to store the sequences
  require("seqinr") # This function requires the SeqinR R package
  choosebank(acnucdb)
  for (i in 1:length(seqnames))
  {
    seqname <- seqnames[i]
    print(paste("Retrieving sequence", seqname, "..."))
    queryname <- "query2"
    query <- paste("AC=", seqname, sep="")
    query2 <- query(`queryname`, `query`)
    seq <- getSequence(query2$req[[1]])
    myseqs[[i]] <- seq
  }
  closebank()
  return(myseqs)
}
```

Part 2 – Decision Trees

- 5) Describe how decision trees could be used to predict gene expression and build gene regulatory networks
- 6) What are strength and weaknesses of RF? Comment on bias of the Gini Index of a split measure
- 7) Given a tree root node, calculate its GINI Index.

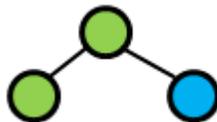
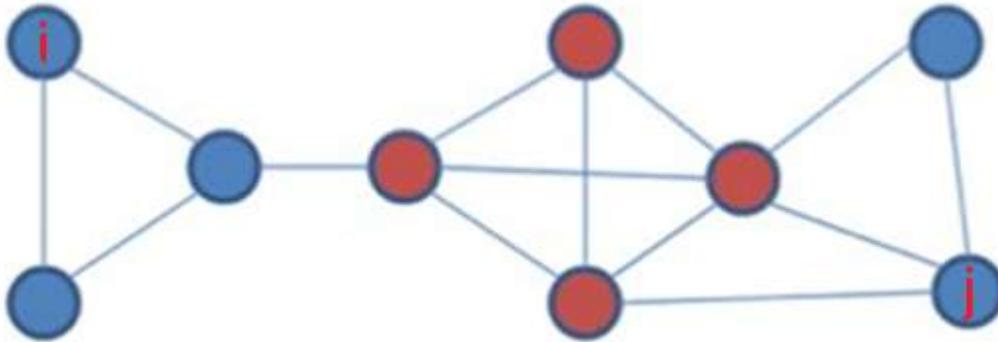


Table 1: Sample values

	Sample 1	Sample 2	Sample 3
class 1	3	6	2
class 2	1	5	9

- 8) What are the two main components of any network?
- 9) Find the shortest path between the node i and j shown in a graph. Use color (red, black) to trace the shortest path



- 10) Define scale-free network and its properties. Give an examples both biological and non-biological networks