Multiple sequence alignment

Exercise 1

Get Sequences from GenBank using R

install and load the following packages:

1. Start with installing and loading the Biostrings library by typing the following commands:

```
> install.packages("ape")
> install.packages("seqinr")
> library(ape) #this is a general R-package for phylogenetics and get the sequence from GenBank
> library(seqinr) #this is an specialized package for nucleotide sequence management
```

2. Using function `read.GenBank` to get the sequence from GenBank.This function connects to the GenBank database, and reads nucleotide sequences using accession numbers given as arguments. Let's read the casque-headed lizard (Basiliscus basiliscus) RAG1 sequence JF80620

```
> seq_1_DNAbin <- read.GenBank("JF806202") #save as DNAbin object
> attr(seq_1_DNAbin, "species") #to get the specie name of the sequence
> seq_1_DNAbin$JF806202
> str(seq_1_DNAbin) # we get the structure of the object
> seq_1_character <- read.GenBank("JF806202", as.character = TRUE) #save as character object
```

3. Extract the sequences for a set of GenBank accession numbers

```
lizards_accession_numbers <- c("JF806202", "HM161150", "FJ356743", "JF806205",
                               "JQ073190", "GU457971", "FJ356741", "JF806207",
                               "JF806210", "AY662592", "AY662591", "FJ356748",
                               "JN112660", "AY662594", "JN112661", "HQ876437",
                               "HQ876434", "AY662590", "FJ356740", "JF806214",
                               "JQ073188", "FJ356749", "JQ073189", "JF806216",
                               "AY662598", "JN112653", "JF806204", "FJ356747",
                               "FJ356744", "HQ876440", "JN112651", "JF806215",
                               "JF806209")
```

4. Get the sequences and save them in a single DNAbin object.

```
> lizards_sequences <- read.GenBank(lizards_accession_numbers) #read sequences and place them in
    a DNAbin object
> lizards_sequences #a brief summary of what is in the object, including base composition
> str(lizards_sequences) #a list of the DNAbin elements with length of the sequences
#notice the one of the attributes is the species names
```

5. It is hard remember which accession number corresponds to which species. So we can use the previous information to create first a vector with such information

```
> lizards_sequences_GenBank_IDs <- paste(attr(lizards_sequences, "species"), names
(lizards_sequences), sep ="_RAG1_")
#build a character vector with the species, GenBank accession numbers, and gene name "_RAG1_ this
     is its common abbreviation: recombination activating protein
```

6. Write sequences to a text file in fasta format using `write.dna()`. However, only accession numbers are included.

```
> write.dna(lizards_sequences, file ="lizard_fasta_1.fasta", format = "fasta", append =
FALSE, nbcol = 6, colsep = "", colw = 10)
#nbcol: a numeric specifying the number of columns per row (6 by default)
#colsep: a character used to separate the columns (a single space by default)
#colw: a numeric specifying the number of nucleotides per column (10 by default).
```

7. Explore the recently created file lizard_`fasta_1.fasta`. This file has the sequences, but only have the accession numbers. We want to rewrite a file to include species information

```
#Read our fasta file using the seqinr package
> lizard_seq_seqinr_format <- read.fasta(file = "lizard_fasta_1.fasta", seqtype = "DNA",
as.string = TRUE, forceDNAtolower = FALSE)

#Rewrite the fasta file using the name vector that was created previously
> write.fasta(sequences = lizard_seq_seqinr_format, names = lizards_sequences_GenBank_IDs,
nbchar = 10, file.out = "lizard_seq_seqinr_format.fasta")
```

Run Clustal Omega

Now we use Clustal Omega to align all the sequences that were extracted from GenBank.

a. Launch Clustal Omega from EMBL-EBL website `http://www.ebi.ac.uk/Tools/msa/`

b. Upload the fasta of DNA sequences that you just retrieved from GenBank. Select DNA from the drop down menu. And then submit your job.

Q1. Click the Phylogenetic Tree tab. Save your Phylogenetic Tree.

Exercise 2

GFP was originally isolated from a jellyfish species called Aequorea victoria. It has since been found in other jellyfish species. The multiple sequence alignment can used used to compare the differences between the GFP from various jellyfish species and GFPuv (the mutated GFP from the pGLO plasmid).

Get Sequences from GenBank directly

Download protein sequences of the following protein, using the access numbers in bold.

AAC53663: GFPuv (cloning vector pGLO is also called pBAD-GFPuv)

AAA27722: green fluorescent protein (Aequorea victoria)

AAN41637: green fluorescent protein (Aequorea coerulescens)

AAK02062: green fluorescent protein (Aequorea macrodactyla)

1. Save the access number into a text file
2. Go to the NCBI website for batch download `http://www.ncbi.nlm.nih.gov/sites/batchentrez`. Upload the file you just created in step 1. Then choose **Protein** for Database and click on **Retrieve**.
3. Click on `Retrieve records for 4 UID(s)`
4. Click on the check box of all sequences. Then click on **Send to** on the top right corner. Choose **File** and select **FASTA**. Click on **Create file** to save protein sequences for these four proteins into a file, called GFP.fasta

Run Clustal Omega

1. Launch Clustal Omega from EMBL-EBL website `http://www.ebi.ac.uk/Tools/msa/`
2. Upload the GFP.fasta, which contains protein sequences. select Protein in the drop down menu. Click on Submit

Note

* - the residues or nucleotides in that column are identical in all sequences

: conserved substitutions have been observed

. semi-conserved substitutions are observed

Q2. View the aligned sequences. Use the accession numbers to identify the sequences.

a. What symbols are used below the alignment when all the amino acids match?

b. What symbols are used below the alignment when the amino acids do not match

Q3. The GFP from A. victoria was mutated to enhance the fluorescence of the protein. The mutated form of the GFP gene was cloned into the pGLO plasmid (pBAD-GFPuv). Identify the amino acids that were changed in the pGLO GFP sequence.

Q4. Click the Phylogenetic Tree tab. Look at the cladogram at the bottom.

a. Which GFP proteins are more alike?

b. Which GFP proteins are most different?

c. Can you make conclusions on the relatedness of these jellyfish species based on this result?

Exercise 3

1. Launch MUSCLE from EMBL-EBL website `http://www.ebi.ac.uk/Tools/msa/`
2. Run multiple sequence alignment of the four GFP protein sequences
3. Click the Phylogenetic Tree tab. Look at the cladogram.

Q5. Are the alignment results from Clustal Omega and MUSCLE consistent?