**Introduction to Pathogen Phylogenetic Analysis**

**Ebola virus genome data: a phylogenetic study**

**Background**

Ebola virus (EBOV), with species name *Zaire ebolavirus*, is a negative-sense single-stand RNA virus of the family *Filoviridae* that causes severe disease in humans. This virus was first described in humans in 1976, from an outbreak that lasted two months and infected 318 people, with a 88% case-fatality rate. Since its description, there have been 12 smaller outbreaks. It is believed that bats are the primary reservoir host. The recent Ebola virus epidemic in West Africa has been the most devastating by far. This epidemic started in at least December 2013 and lasted until 2016, with a total of 28,646 confirmed cases and 11,323 deaths (Holmes et al. 2016). Evolutionary studies revealed important insights about this outbreak. For example, phylogenetic methods were used to infer when and where the outbreak started, and how quickly it had been spreading. Importantly, such insight was possible as early as 2014 due to the recent surge of large-scale methods to generate real-time genomic data.

In this prac we will analyse a data set of EBOV sequences. It includes historical samples from 1976, and it will be useful to estimate the evolutionary rate and time of origin of the virus.

**Practical 1a: Sequence alignment in MEGA**

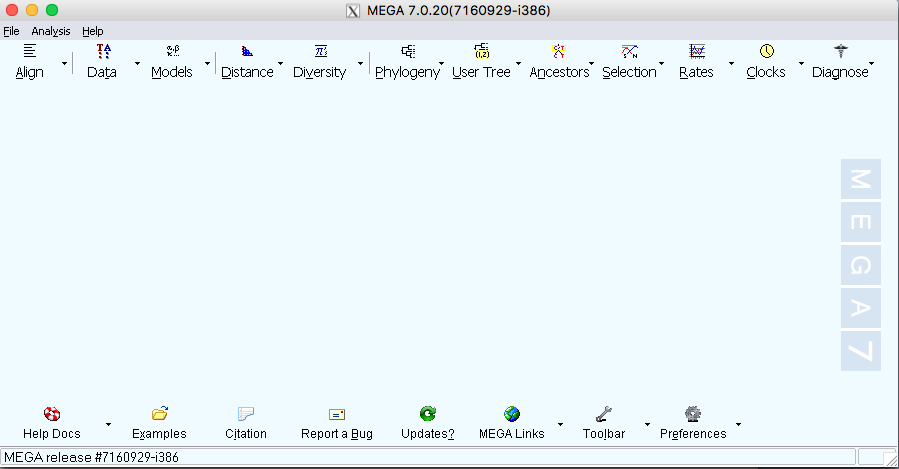
**Data set**

* Sequence data in fasta format of samples collected before the 2013-2016 epidemic: EBOV\_N2\_unaligned\_early\_samples.fasta

**Software**

* MEGA v7 (megasoftware.net)

Open MEGA v7 in your computer, you should see this window appear (Fig 1):



**Fig 1.** MEGA graphical interface

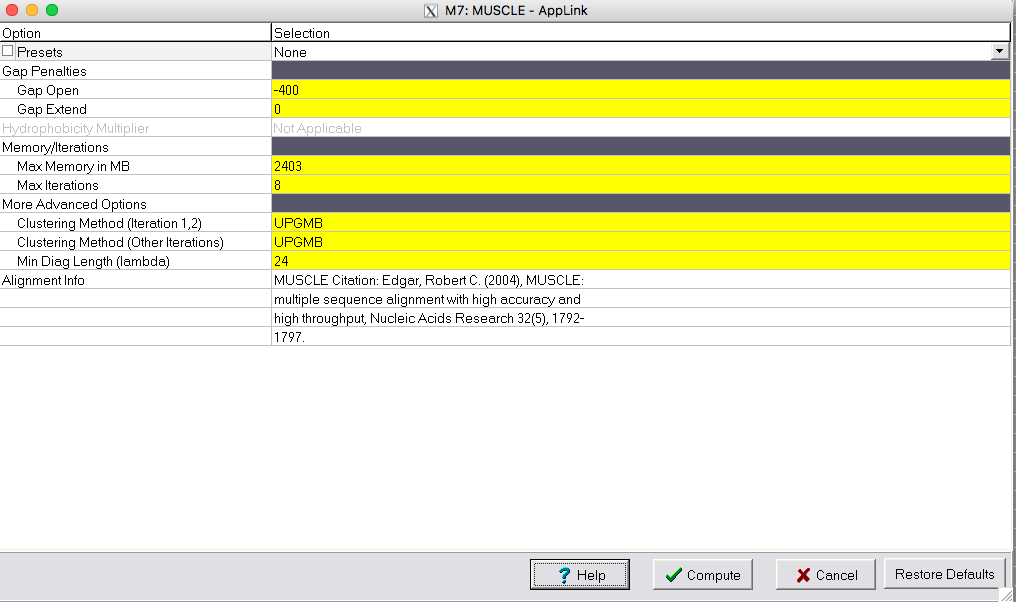
Click on *Data* and then on *Open A File/Session…* Find the file EBOV\_N2\_unaligned\_early\_samples.fasta. A window will come up, asking you whether you want to analyse the data or align it, select *Align*:



**Fig 2.** Ebola virus nucleotide sequence data in MEGA.

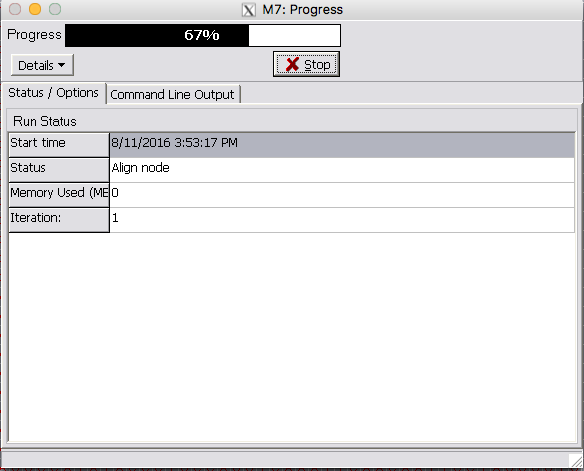
You should see the sequence data, as shown in Fig 2. Each row corresponds to a sequence, and the columns are the different sites. Note that individual sites do not appear aligned. That is, we could not consider them homologous. To align them, click on the symbol and select *Align DNA*. This will conduct a sequence alignment using MUSCLE. If a window pops up, telling you whether it should select all the sequences, click *OK*.

Next, a window with alignment options will appear (Fig 3).



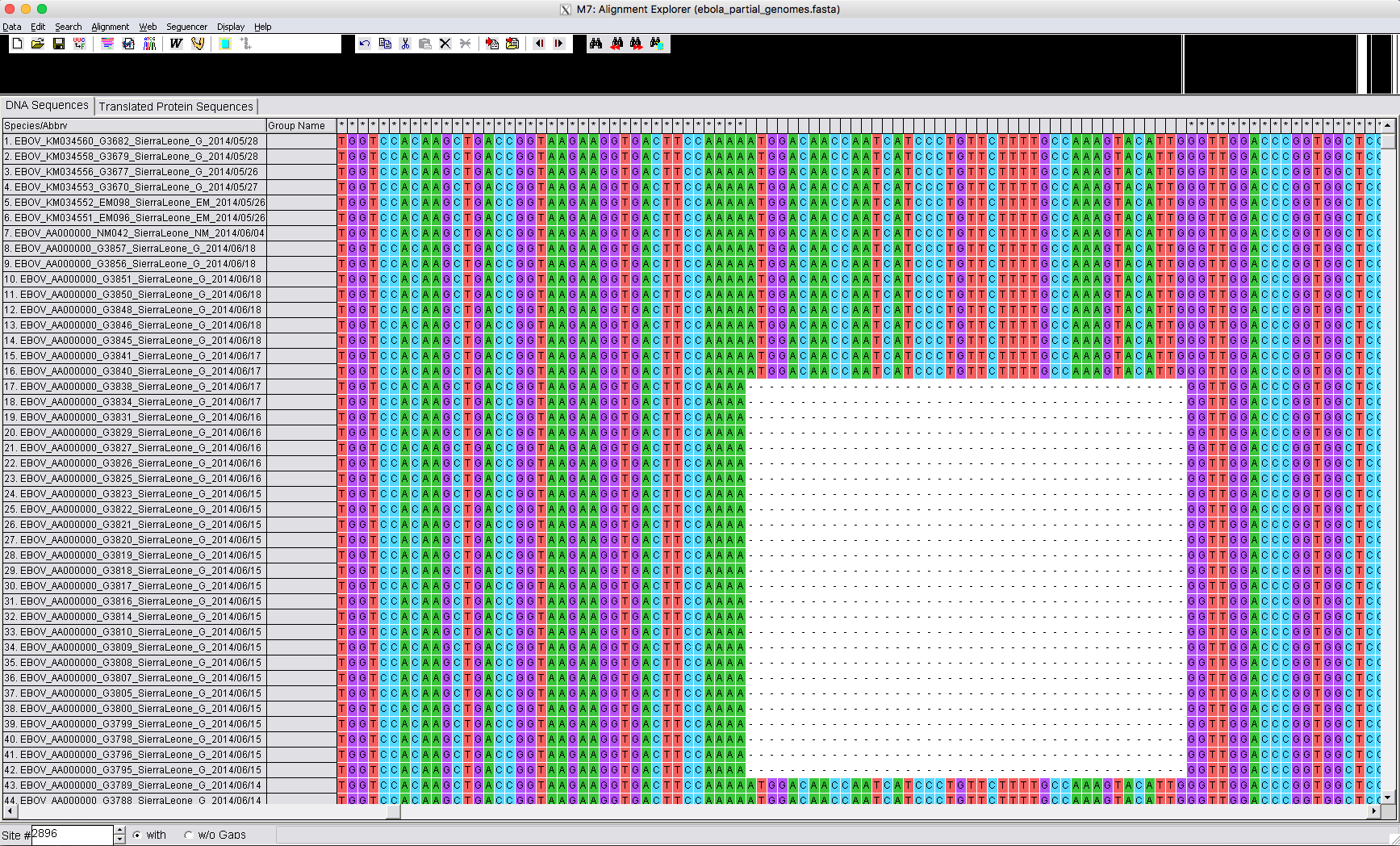
**Fig 3.** Alignment options for the MUSCLE algorithm.

Click *Compute*. The default options are fine for this analysis. In practice you can compare the results between different alignment options. For example, if the resulting alignment has too many gaps, you can increase the penalties for opening and extending gaps. You will see a window appear (Fig 4), which shows the progress of the alignment iterations.



**Fig 4.** Progress of sequence alignment in MEGA.

To see more details, click on *Command Line Output*. The analysis will run for a few minutes. Once it is complete, scroll horizontally to see the resulting alignment. Note that there are a few gaps, but the sequences now look aligned (Fig 5).



**Fig 5.** Sequence alignment in MEGA with gaps.

We want to save the alignment. To do this, click on *Data*, and on *Export* *Alignment*. Select FASTA format and name it: EBOV\_N2\_aligned\_early\_samples.fasta

Repeat this step to save the file in Phylip format, and name it:

EBOV\_N2\_aligned\_early\_samples.phy

Close MEGA, do not save the working session - we only need the alignment that we saved in the previous step.

Open the alignment saved in fasta and phylip format in a text editor and get some familiarity with how they differ. Some phylogenetics programs nowadays accept many common sequence alignment formats, but some programs, like PhyML, are very specific about the format they accept.