**Practical 1b: Model selection in MEGA**

**Data set**

* Sequence alignment in fasta format of samples collected before the 2013-2016 epidemic:

EBOV\_N2\_aligned\_early\_samples.fasta

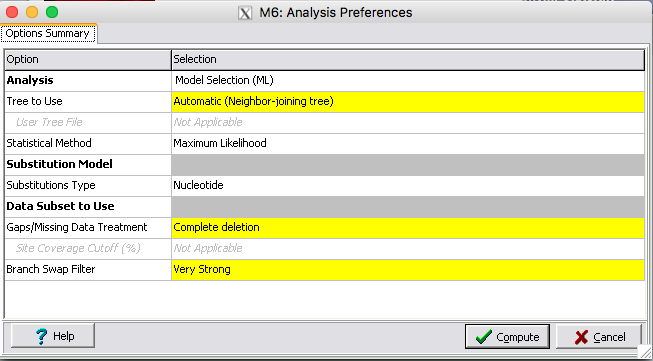
**Software**

* MEGA v7 (megasoftware.net)

Open MEGA and click on *Data*, *Open A File/Session*, and find the alignment from the previous prac (EBOV\_N2\_aligned\_early\_samples.fasta). The program will ask you whether you want to analyse or align the data. Select *Analyze*.

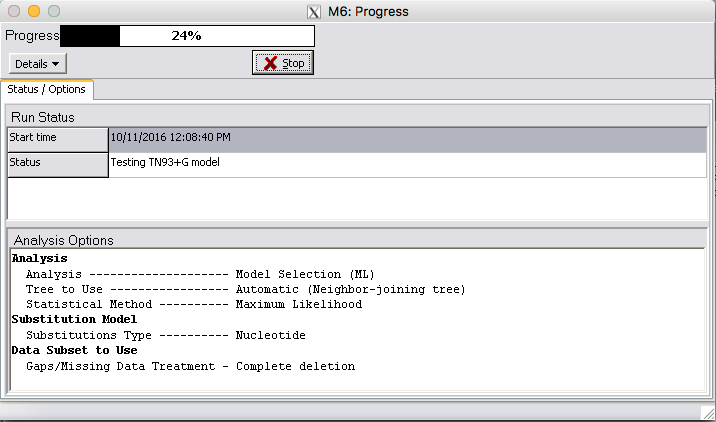
MEA will then ask you whether the data correspond to nucleotide or protein sequences. Select *Nucleotide Sequences* and click OK. It will also ask whether the data are protein coding sequences. Click on *No*. Although they correspond to a protein-coding gene, codon models are very computationally intensive, so we will use nucleotide models instead.

Once the data are loaded in MEGA, click the  button and select *Find Best DNA/Protein Models (ML)…* If it asks you whether you want to use the active data, select Yes. The window in Fig 1 will appear.



**Fig 1.** Substitution model selection menu.

The default settings for the model selection procedure are fine. Click on *Compute*. A window with the progress of the analysis will appear (Fig 2).



**Fig 2.** Progress of model selection in MEGA.

MEGA performs a very thorough model selection. It tests 24 substitution models and calculates the likelihood, and the AICc and BIC scores.

**Question 2.1:** What is the optimal substitution model for these data? What assumptions does it make about the evolution of these sequences?

**Question 2.1:** How many parameters does the optimal model have? What do these parameters represent?

**Question 2.3:** Do the BIC and AICc agree on the optimal model chosen?