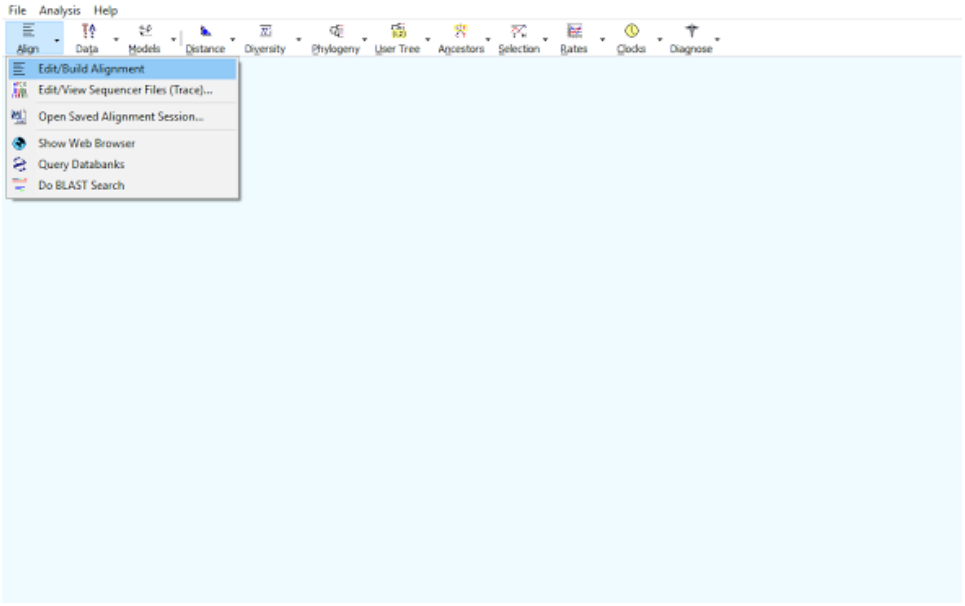
Build a phylogenetic tree using MEGA7 (Molecular Evolutionary Genetics Analysis Version7)

Before starting:

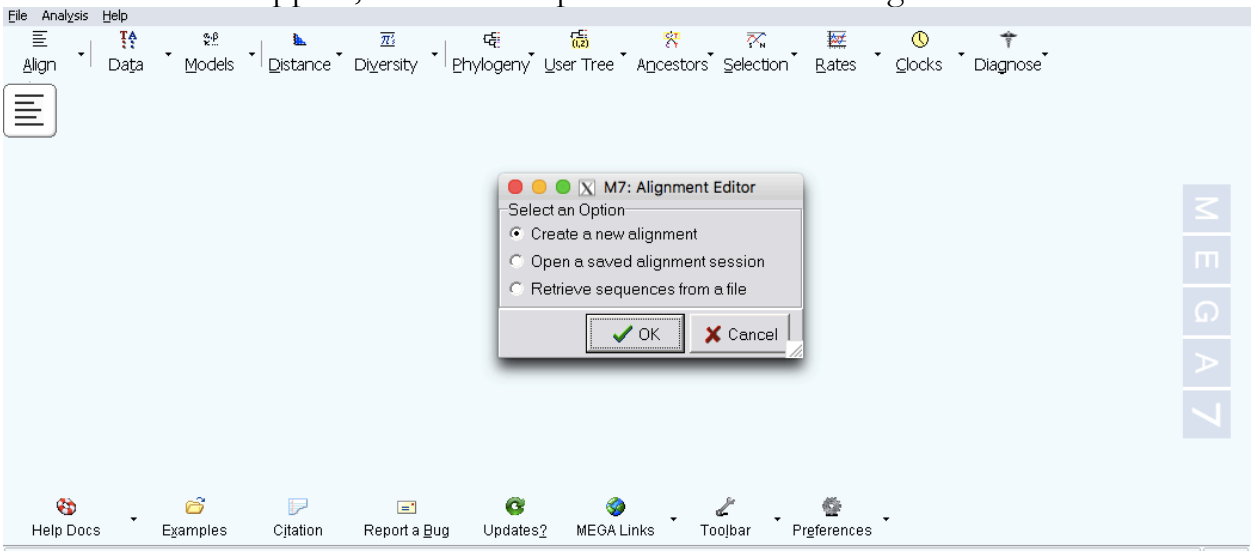
1. In case you still don't have MEGA7 on your computer, download the graphical version of the program at http://www.megasoftware.net/ and install it.
2. Make sure you have downloaded the sequence files in a fasta extension. You will need at least 3 files for each scenario: the reference sequences, the outgroup and the file(s) with the sequences to study

Step one: load the sequences in MEGA7

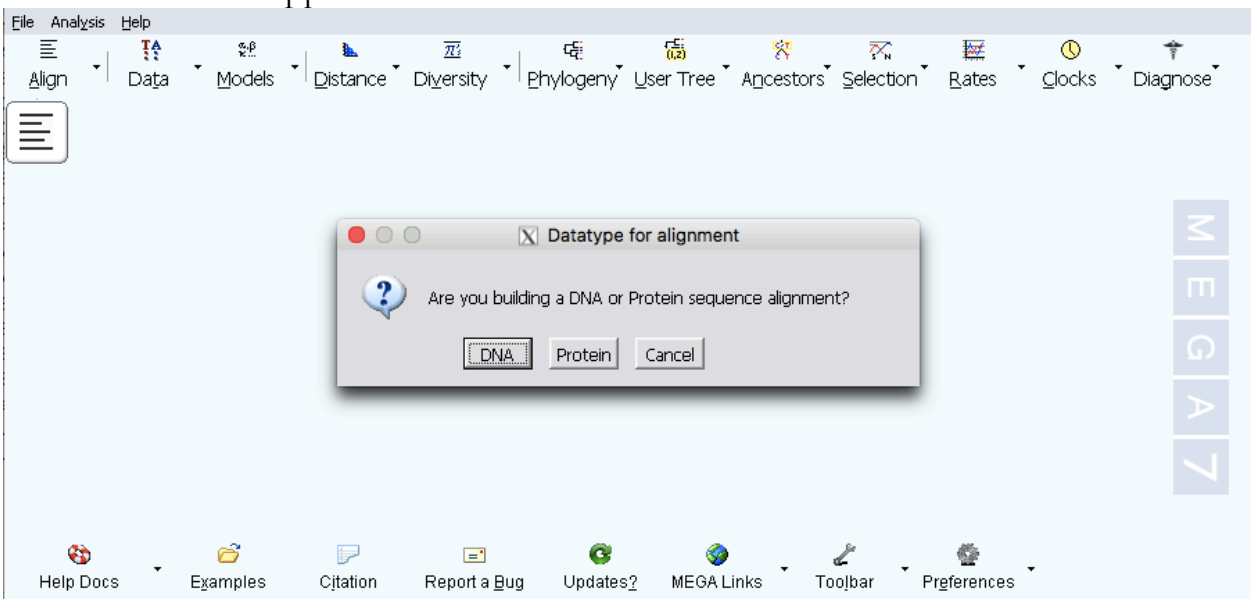
1.0 Select the option "Align" and choose "Edit/build Alignment" in the top left of the toolbar in MEGA7



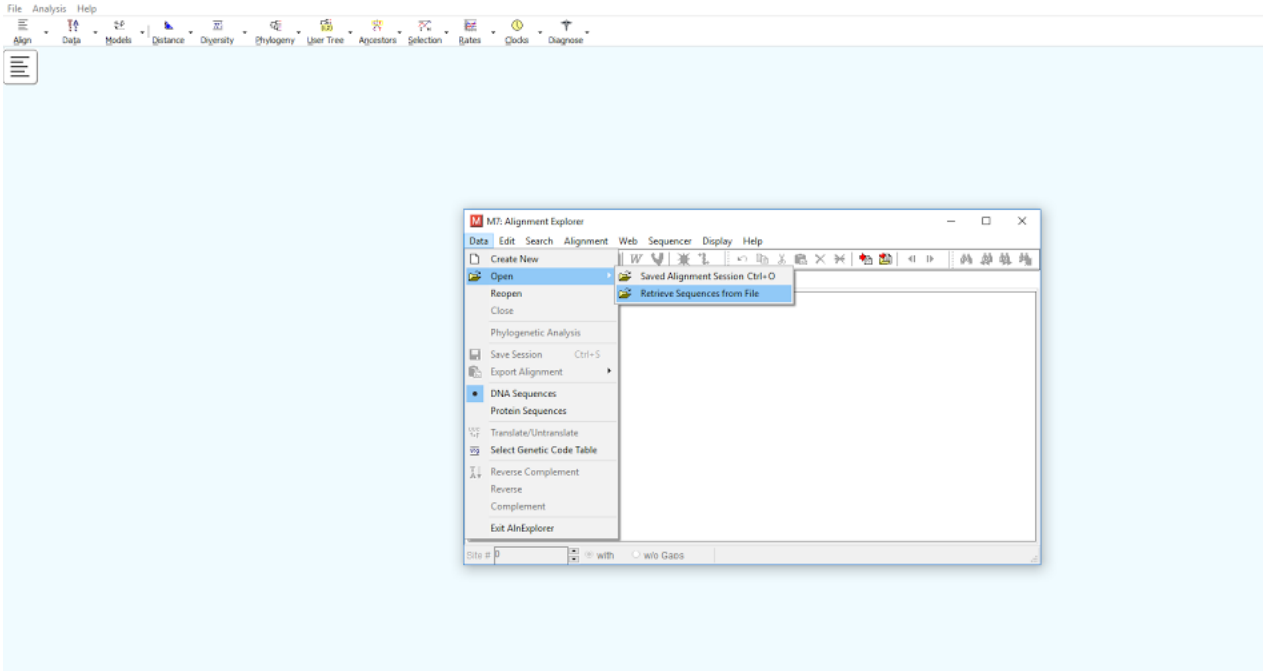
1.1 A small box will appear, choose the option "Create a new alignment"



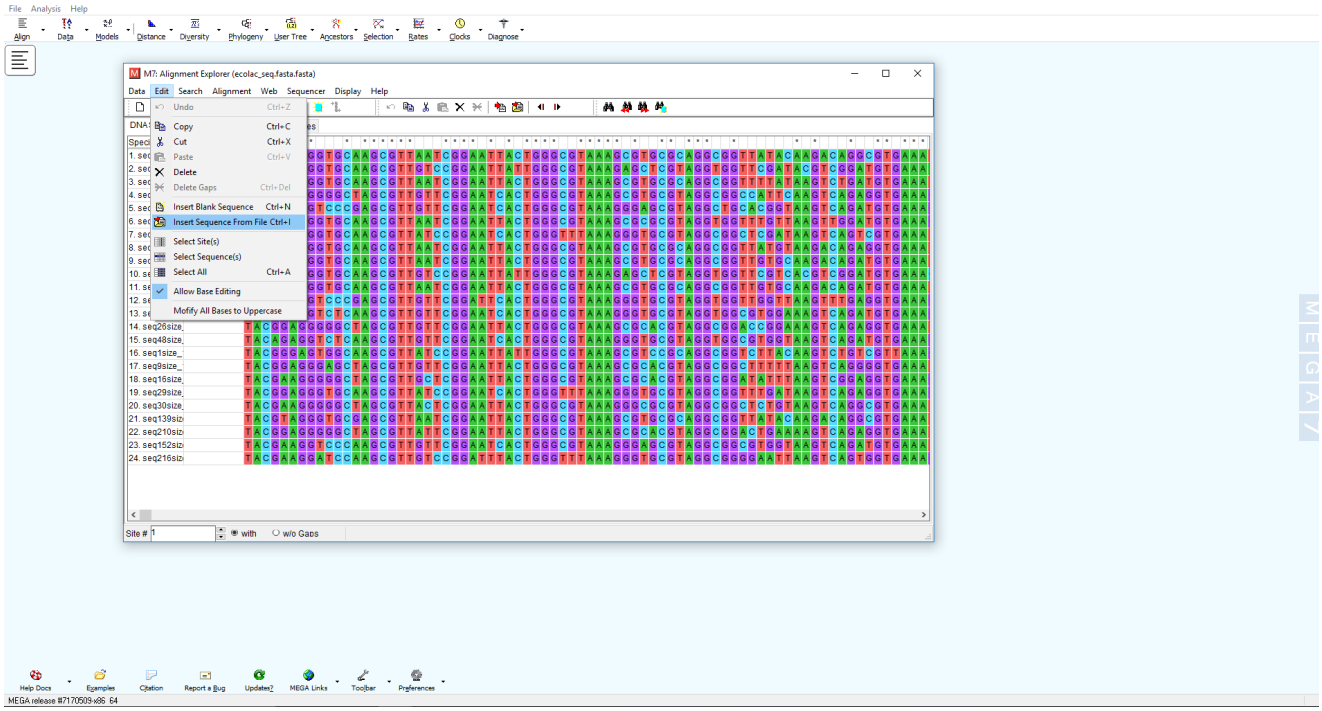
1.2 Another box will appear: select “DNA”



1.3 Now go to “Data → Open → Retrieve sequences from file” and choose the fasta file with the sequences you want to study.

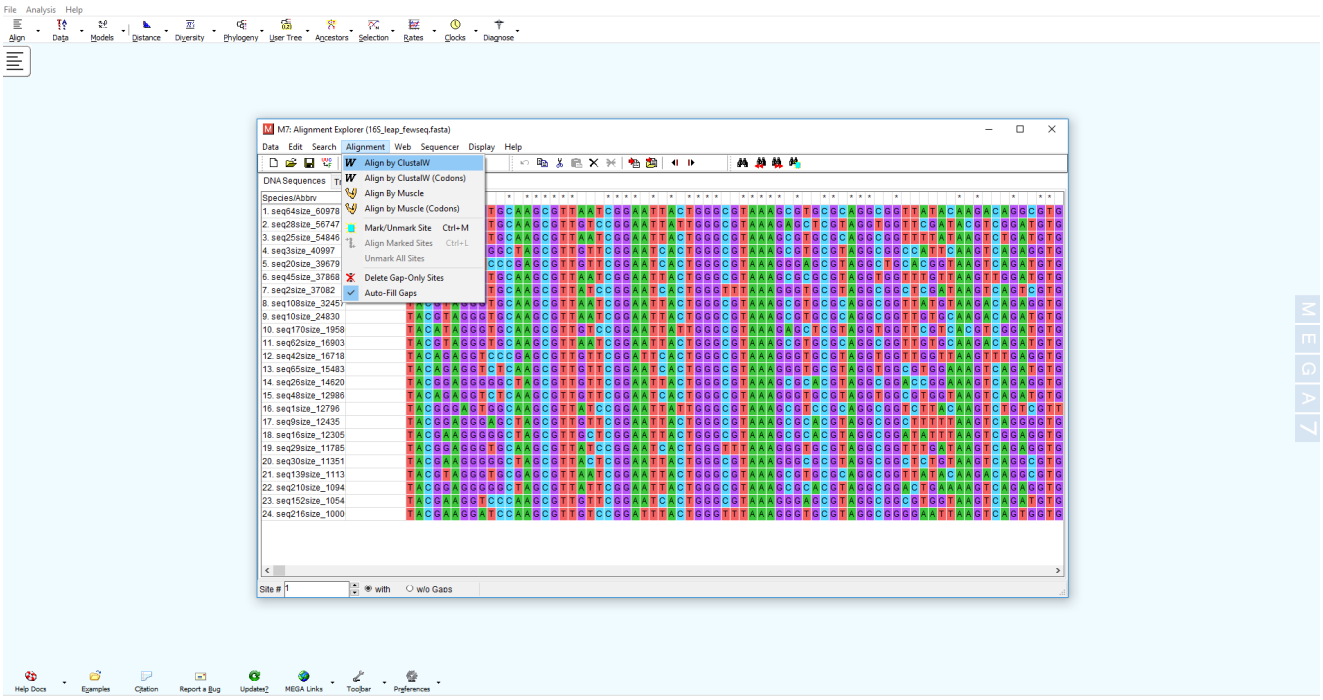


1.4 To add the other sequences that you want to align go to “Edit => Insert Sequence From File”. Don’t forget to add the outgroup, which will be important while placing the root of the tree, and the reference sequences that help to have an idea of the non-cultured sequences identification when the tree is done.

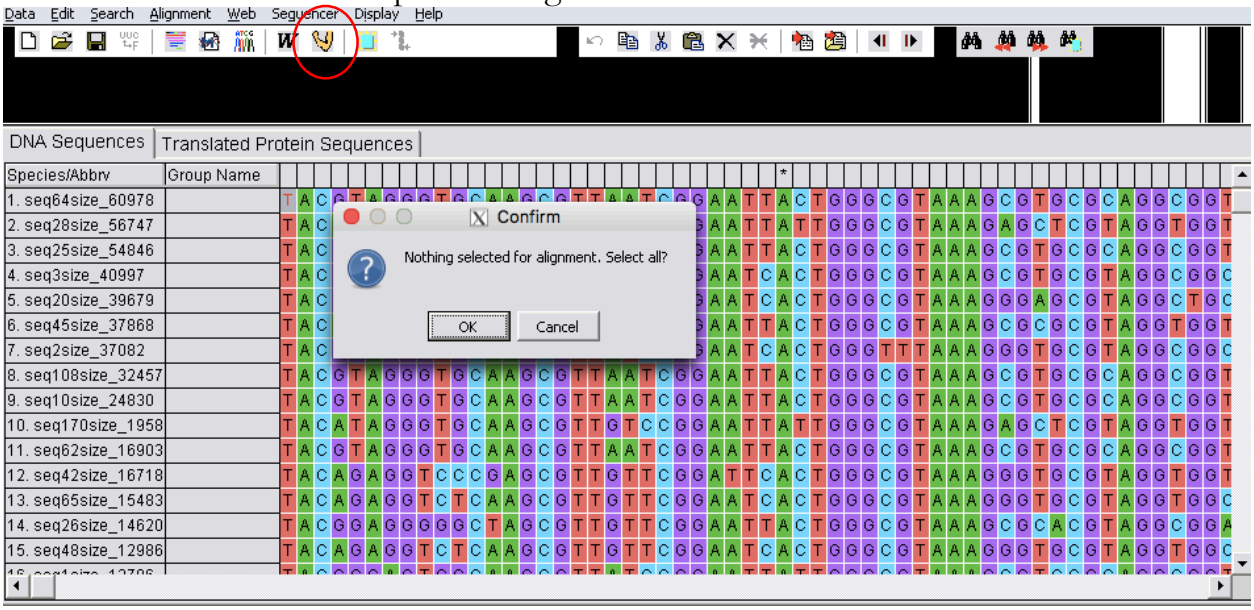


Step two: perform the alignment

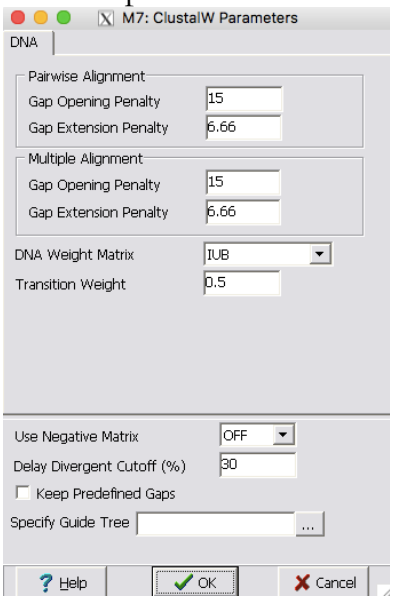
2.1 Click anywhere outside the first column to avoid selecting only a row of sequences to align. Select the option "Alignment" and choose one method to align the sequences, could either be by ClustalW or Muscle. Here we used ClustalW.



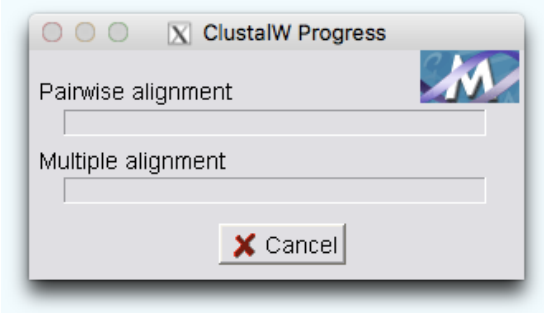
2.2 Press OK to select all samples for alignment



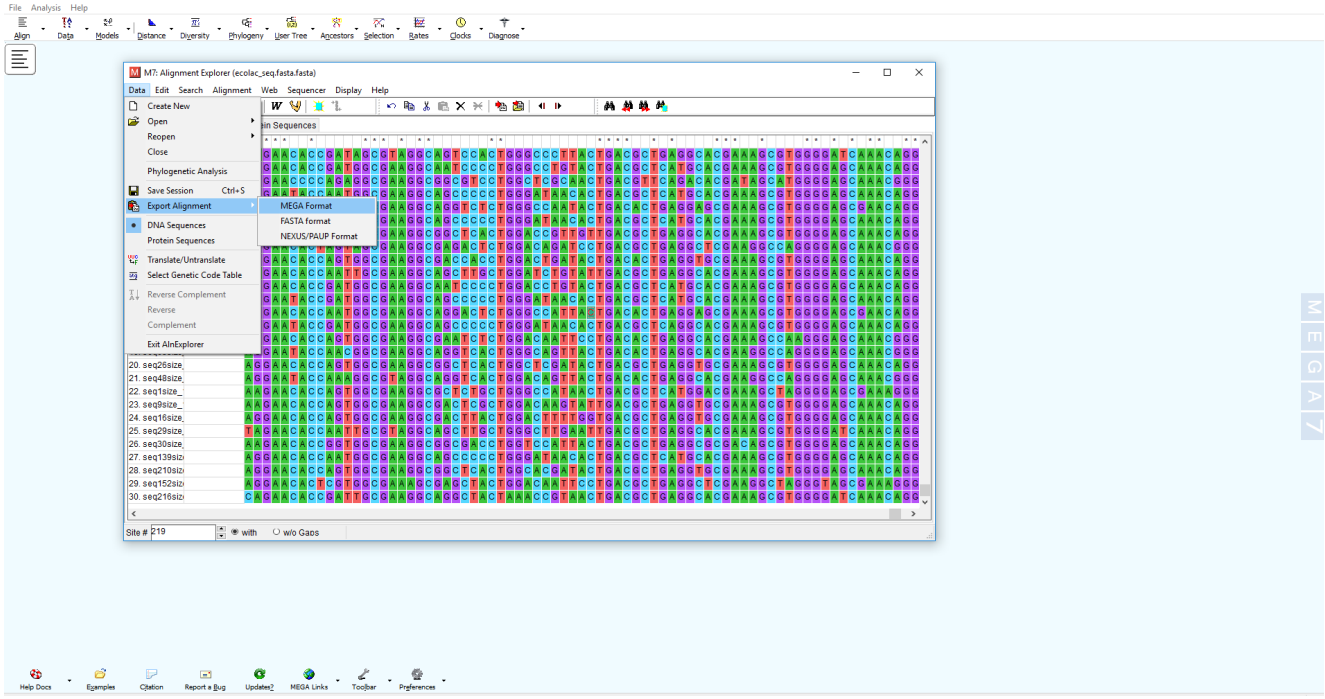
2.3 Press OK again to accept the default parameters for the alignment



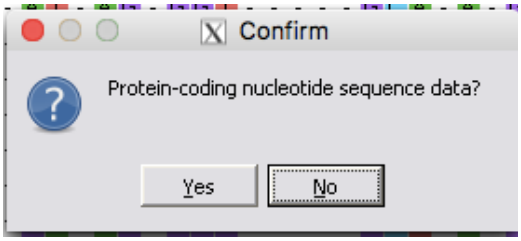
2.4 And wait until the alignment is done (it will take more time the more sequences you have and the longer they are). A box like this will show the progress of the alignment.



2.5 Once the alignment is done you can export the alignment in a MEGA or FASTA format by clicking on "Data → Export alignment" and selecting these format options in the toolbar.

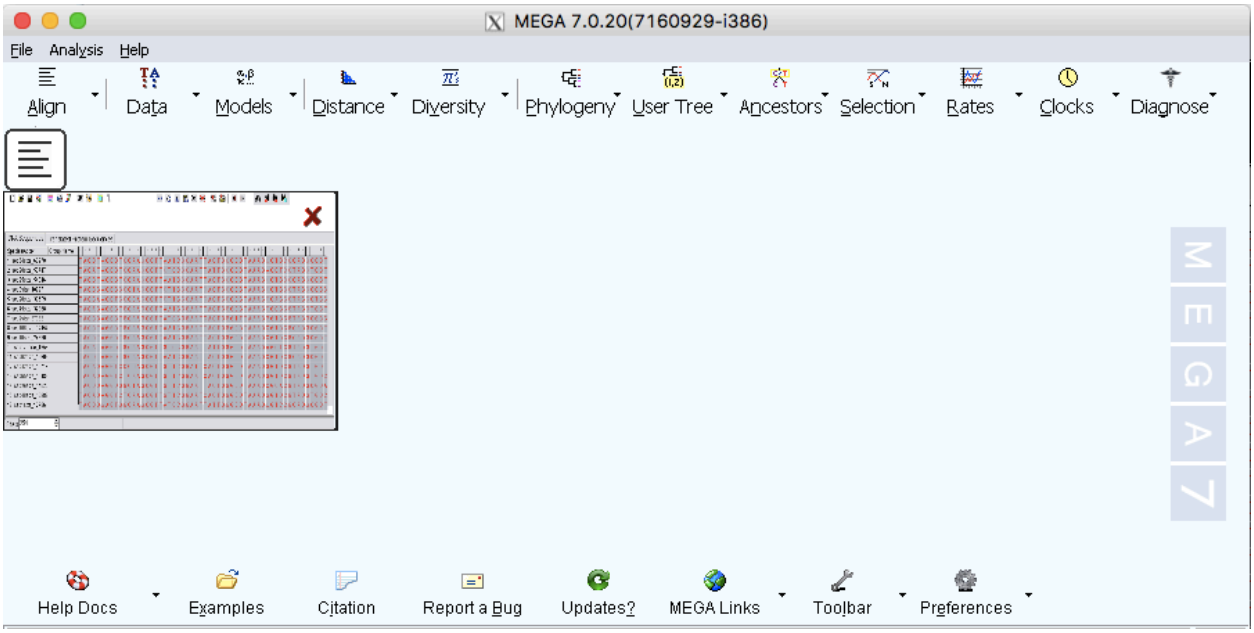


2.6 To save your alignment, go to "Data → Export Alignment → MEGA format". Choose where to save it in your computer, create a title for your data and confirm that the sequences are NOT protein-coding sequences.

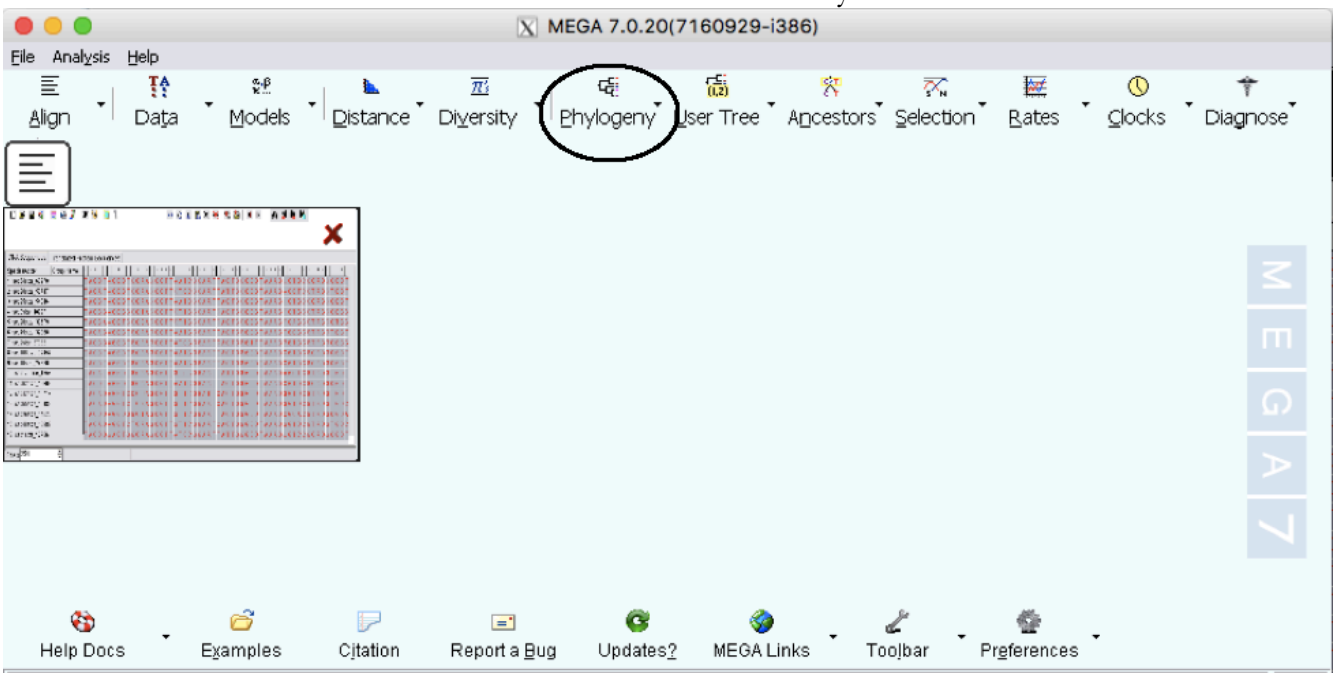


Step three: build the phylogenetic tree

3.0 Go back to the main screen on MEGA and if you haven't close your alignment session you will see it there by clicking on the alignment sign 

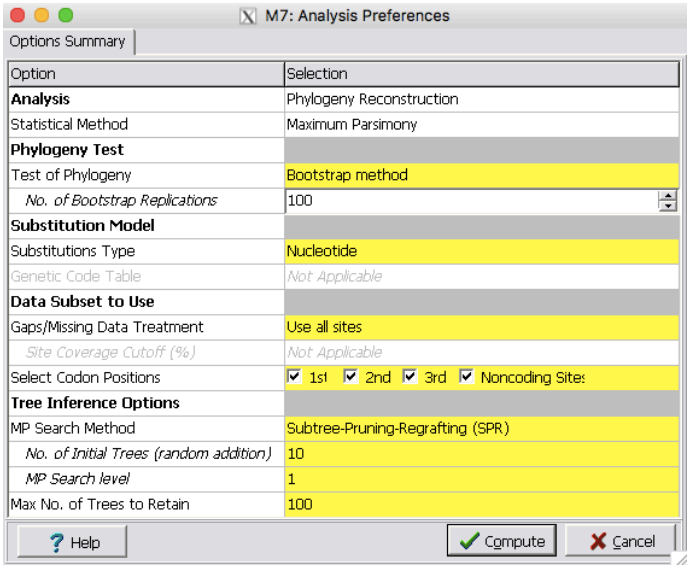


3.1 Select the option "Phylogeny" in the tool bar and choose the method you want to build your phylogenetic tree. In the end of this document there is a brief explanation about these different methods. Here we used the Maximum Parsimony Tree.

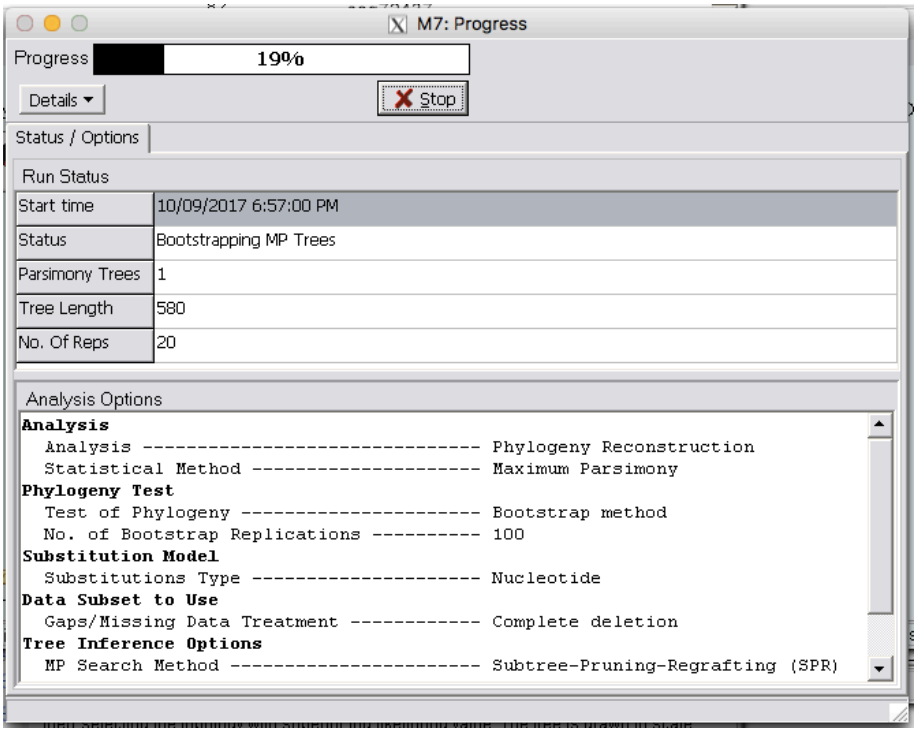


3.2 The program will ask you to choose the file (in a FASTA or MEGA format) to analyse. Load the file you saved in the step 2.6

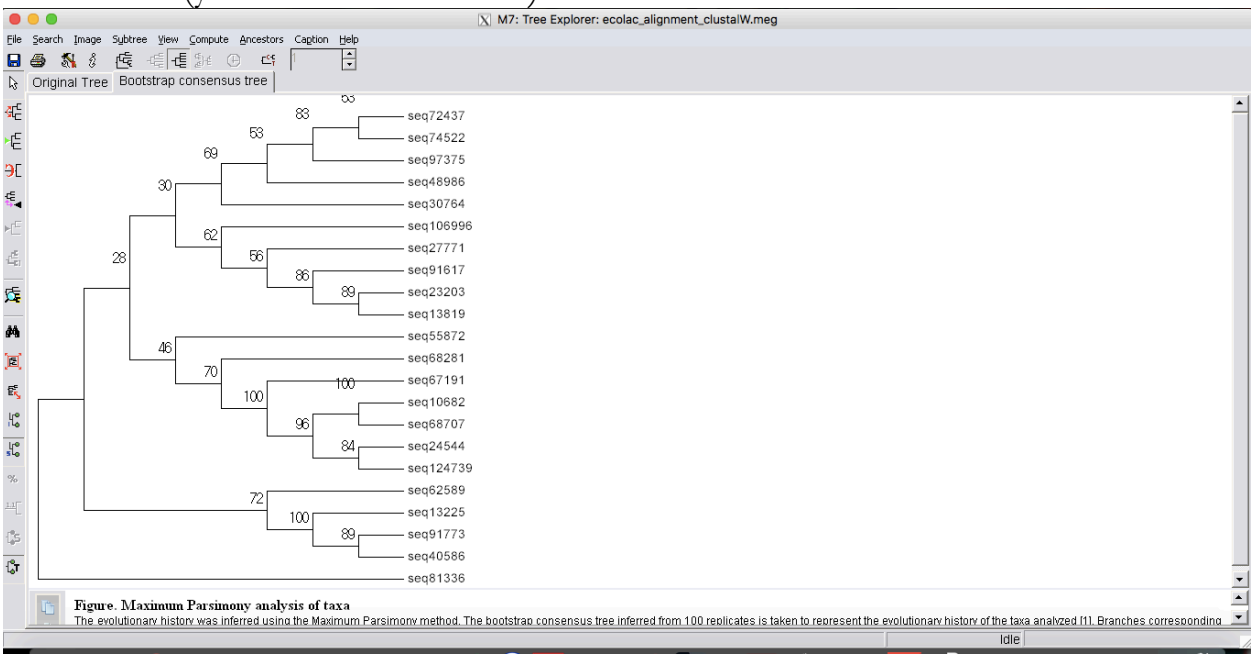
3.3 After loading the data, you will have access to the analysis preferences box of the method you chose. Select "use all sites" in "Gaps/missing data treatment". In the "Test of Phylogeny" option choose the bootstrap method and in "No. of bootstrap replications" write "100". By choosing these options the program will run 100 replications of a phylogenetic tree based on the sequences you loaded and the parameters you chose and return the most probable tree with bootstrap values for each branch. We chose 100 replications for a matter a time, but the more replications, the more confidence we have in our tree. You can keep the other parameters as default.



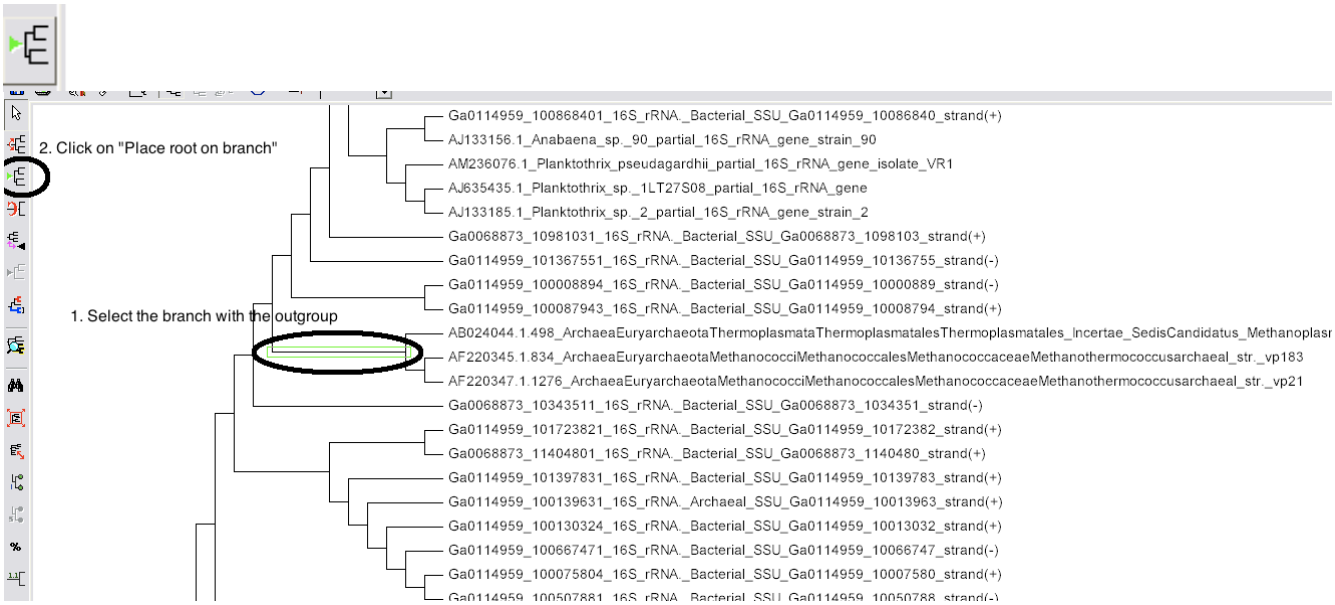
3.4 Wait until the analysis is done. You can check in the progress bar how it is going and in the "Analysis options" the summary of parameters chosen for it. It will take a while until it's done, go grab a coffee!



3.5 Once the tree is done, the program will show you the Original Tree and the Bootstrap consensus tree (you should use this last).



3.6 Congrats! You have built a phylogenetic tree based on 16S rRNA gene sequences! Now you can format it using the toolbar in the left and answer your research question! Don't forget to place the root: select the branch where you find the outgroup added and then click on the bottom "Place root on branch". It is the third bottom on the toolbar on the left and it looks like this:



Step four: save and export the phylogenetic tree

4.1 You can save your phylogenetic tree as a PDF by clicking on "Image → Save as PDF file" or also save in a MEGA format (to be able to open it in the program again) by going on "File → Save current session". For the next step of the workshop, save your tree as a newick file (.nwk) fo to "File → Export as NEWICK").

Glossary of terms

- A **FASTA file** or a file in a FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format allows for sequence names and comments (that follow the sign ">") to precede the sequences.
Example of a FASTA format of a DNA sequence:
>sequence_1
TACGAAGGGGGCTAGCGTTACTCGGAATTACTGGGCGTAAAGGGCGCGTAG
GCGGCTCTGTAAGTCGGCGTGAAATTCCTGGGCTCAACCTGGGGGCTGCGC
TTGAGACTGTGGGGCTAGAGGATGGAAGAGGGTCGTGGAATTCCCAGTGTA
GAGGTGAAATTCGTAGATATTGGGAAGAACACCGGTGGCGAAGGCGGCGAC
CTGGTCCATTACT

- A **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Very short or very similar sequences can be aligned by hand. However, most interesting problems require the alignment of lengthy, highly variable or extremely numerous sequences that cannot be aligned solely by human effort. For this reason, different algorithms, such as the ClustalW and Muscle which are present on MEGA7, perform sequence alignments.

- A **phylogenetic tree**, also known as a phylogeny, is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor (Baum 2008). There are different methods for phylogeny estimation. The Neighbor-Joining method, for example, is based on a distance matrix of all possible pair of taxa. There are also methods based on the parsimony approach, which purpose's is to find a phylogeny that requires the fewest necessary changes to explain the differences among sequences. Finally, the maximum likelihood method assigns probabilities to different trees and then pick the tree with the most probable one (i.e. the one with highest likelihood values) (source: NCBI Advanced Workshop).

References

- Baum, D. 2008 Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups. Nature Education 1(1):190
- Kumar et al. 2016 MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016 Jul. 33(7):1870-4
- Wikipedia:
https://en.wikipedia.org/wiki/Sequence_alignment and
https://en.wikipedia.org/wiki/FASTA_format
- NCBI Advanced Workshop for Bioinformatics Information Specialists, Module title: Phylogenetic resources. Available on
https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Phylogenetics/phylo12.html