

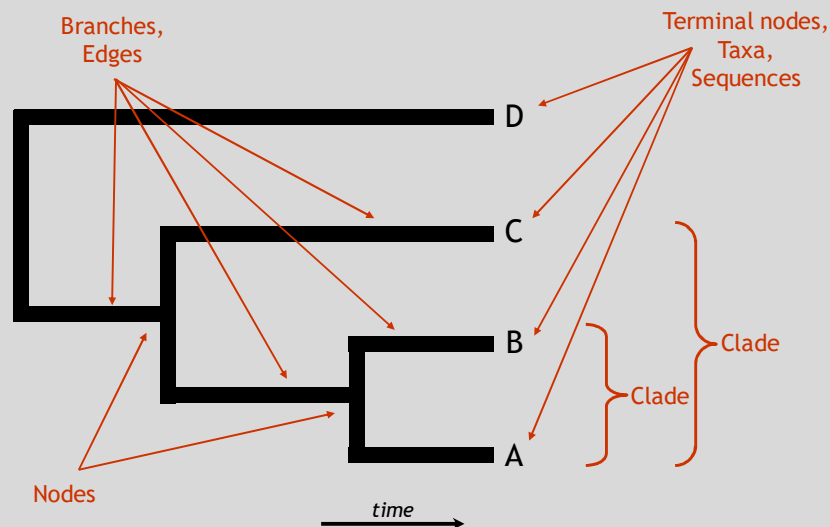
Lab 4: Phylogenetics

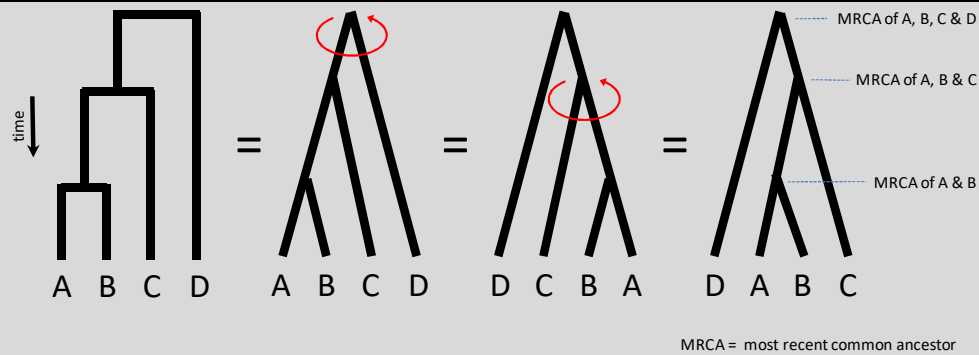
[Software needed: MEGA X, and web access]

Today we'll use the multiple sequence alignment (MSA) we generated from the last lab to infer the evolutionary relationships between the sequences. Phylogenetic analyses produce branching diagrams that can clearly illustrate relationships between sequences that are not apparent from analyses such as BLAST or MSAs. Phylogenetic trees are obviously useful for evolutionary and comparative studies focused on evolutionary relationships and patterns of divergence, but they are also becoming increasingly important for generating hypotheses regarding gene or protein function for molecular and biochemical studies.

Phylogenetics is a huge field, and frequently by itself takes up an entire course. We will simply skim the surface in this lab by introducing some very basic tools and ideas. There are two primary categories that most phylogenetic tools fall in: distance-based methods, and character-based methods. We will use one of each of these approaches. Your goal in this lab is not only to get a feel for how to build a phylogenetic tree, but also to gain a greater appreciation for the fact that these are not cut-and-paste analyses, as is too often assumed for bioinformatic approaches. Often times, different approaches can provide different conclusions; consequently, you need to play with multiple methods and parameter values, and ultimately use your budding biological intuition to generate the best tree and the strongest analysis. See **Box 7** for how such analyses can be interpreted.

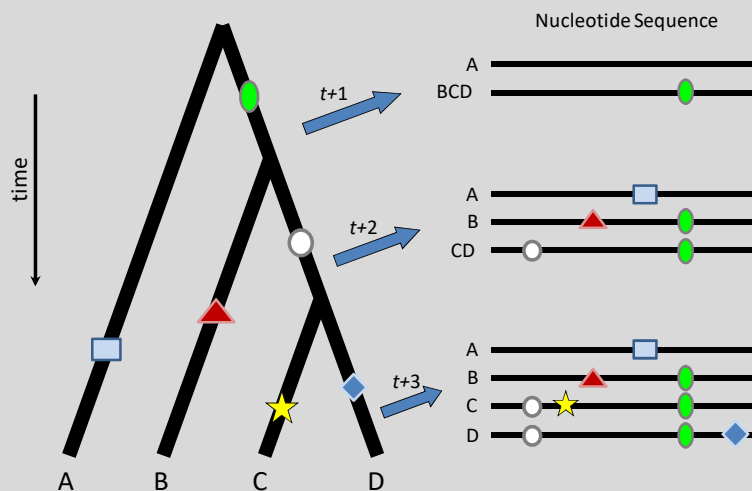
Box 1. Anatomy of Phylogenetic Trees





Phylogenetic trees can be presented in many different shapes and orientations. The important thing to recognize is that the only way to determine the evolutionary distance between two sequences is how far back in time you have to go before finding a common ancestor. So, in the tree on the right, even though *sequence A* is closer to *D* than to *C* physically on the page as letters, it is in fact more closely related to *C* since they share a more recent common ancestor. The evolutionary relationships between A-D can also be represented using **Newick** format as follows (((A, B), C), D): the nesting of the parentheses follows the bifurcation of the trees above.

Box 2. The Growth of Phylogenetic Trees



As organisms evolve and diversify, their lineages accumulate mutations (represented by the coloured shapes). These mutations will be transmitted to all descendent offspring and lineages, so a mutation that occurs very early in the history of the group, e.g. the green oval mutation, which occurred in the ancestor of sequences B, C & D, will be found in all three descendent lineages. A mutation that occurs later (e.g. the blue square mutation) is likely to be found in fewer lineages. The position of the mutations on the nucleotide sequences is completely arbitrary and meant only to show how many unique sequences there are at each point in time, and the distribution of mutation among these sequences.

Neighbour-Joining

Box 3. Distance-Based Methods

While there are very substantial differences between how these methods work, all distance based methods approach phylogenetic reconstruction by first generating a ‘distance matrix’ by performing pairwise comparisons between all sequences, and calculating the genetic distance between each pair. In the simplest case, the genetic distance is just the number of mismatches between the sequences, although most distance matrices use more sophisticated distance measures. Once again, in the simplest case, the pairwise distance matrix is then used to identify the two most closely related sequences, which form the first two branches of the tree. These pairwise distance matrix is then remade, but this time the two most similar sequences that were identified in the last step are represented by a single node in the tree. Now the next most similar pair of sequences (or sequence - node, or node - node) are identified, and the process is repeated until the table collapses to as single node. All the distances between the sequences and the nodes are then used to draw a tree.

Distance-based methods have the advantage of being very fast to compute and simple to construct. The trees produced by methods such as Neighbour-Joining or Minimum Evolution are also quite reliable for most datasets. Nevertheless, all distance-based methods also suffer from one or more fairly strong underlying assumptions. The most important of these requires what are called ‘additive distances’ – the details of which are beyond the scope of this course (see the Further Readings for more details if you’re interested). Suffice it to say that this assumption may be violated if a subset of your sequences has evolved at a much fast rate than others. Violation of this assumption can seriously compromise your analysis.

We’ll start by building a neighbour-joining tree in MEGA X. You were introduced to MEGA last lab when you performed MSAs. MEGA (Molecular Evolutionary Genetic Analysis) is a very easy to use, yet powerful application primarily used for phylogenetics. Neighbour-joining is fast and quite reliable, and therefore makes great starter trees to build hypotheses about general tree topology/distance, before moving on to more rigorous programs.

1. Open MEGA X and convert the aligned DNA file downloaded from the “Reading: Lab 4 -- Phylogenetics” section from FASTA format to MEGA format. (The file is “*bioinformethods1%2Flabs%2FLab3,4_sequences_DNA_aligned.fas*”; we aligned these sequences last lab but we’re providing this alignment so everyone’s starting with the same alignment. Get it from the part of the Coursera site where you downloaded this lab manual; right- or Command-click on the [aligned DNA](#) link there to download the alignment in Fasta-format to your computer for use here. If you’re a Mac user, make sure to remove the .txt extension that your Mac will add to the filename, or you won’t see it in MEGA).
 - Click **File/Convert File Format to MEGA...**(see Figure 1)
 - To locate the file you need to click on the very small folder icon on the right side of the **Data file to convert** box.
 - Locate your ***aligned nucleotide sequence file*** in FASTA format. Select the Data Format (FASTA) if is not automatically recognized.

- Click **OK**
- Assuming the file converted correctly, save it with an informative name. You will notice that the file is appended with a **.meg** file extension. MEGA and FASTA format are very similar for these simple files, the major difference is that MEGA format can store more information in different fields.
- MEGA format can be a little picky. Here are the rules:
 - Input files cannot have sequence names with blank spaces, or any of the following characters
 $, ; : ' " ! ? > < [] \sim @ \# \wedge \&$
 - Curly brackets can only be used if they are paired.
 - The first line is always: **#MEGA**
 - The second line is always: **!Title: xxx**, where xxx is anything

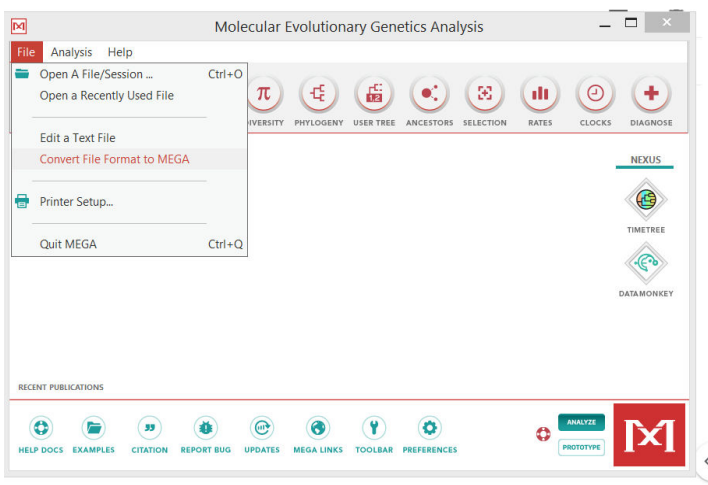



Figure 1. MEGA X user interface.

2. Open up the new file
 - Click **File > Open a File/Session**; that is the first menu option in **Figure 1** (make sure the Text File Editor and Format Converter Window is closed)
 - Locate and open the new file with the **.meg** extension.
 - Specify that this file contains **Nucleotide Sequences**, and any other information requested (protein coding sequence = Y, select genetic code = standard).
 - Click on the “TA” icon to open the Data Explorer. This is useful for visualizing and selecting data and regions for analysis.
 - Note: if there is an illegal character in your file, MEGA will tell you which line it is located on.
3. Go back to the main window, and select **Analysis > Phylogeny > Construct/Test Neighbor-Joining Tree**. We'll stick with the default parameters this time but one: make sure to choose “Bootstrap method” under **Phylogeny Test/Test of Phylogeny**. Click **Compute**.

MX: Analysis Preferences

Phylogeny Reconstruction	
Option	Setting
ANALYSIS	
Scope	→ All Selected Taxa
Statistical Method	→ Neighbor-joining
PHYLOGENY TEST	
Test of Phylogeny	→ Bootstrap method
No. of Bootstrap Replications	→ 500
SUBSTITUTION MODEL	
Substitutions Type	→ Nucleotide
Genetic Code Table	→ Not Applicable
Model/Method	→ Maximum Composite Likelihood
Fixed Transition/Transversion Ratio	→ Not Applicable
Substitutions to Include	→ d: Transitions + Transversions
RATES AND PATTERNS	
Rates among Sites	→ Uniform Rates
Gamma Parameter	→ Not Applicable
Pattern among Lineages	→ Same (Homogeneous)
DATA SUBSET TO USE	
Gaps/Missing Data Treatment	→ Complete deletion
Site Coverage Cutoff (%)	→ Not Applicable
Select Codon Positions	→ <input checked="" type="checkbox"/> 1st <input checked="" type="checkbox"/> 2nd <input checked="" type="checkbox"/> 3rd <input checked="" type="checkbox"/> Nonc
SYSTEM RESOURCE USAGE	
Number of Threads	→ 3

Figure 2. MEGA X Neighbour-Joining analysis window, with nucleotide sequence defaults (except that a Bootstrap method Test of Phylogeny has been selected). Ensure “Complete Deletion” is selected for Gaps/Missing Data Treatment.

4. Your output should be a nicely formatted phylogenetic tree in a new window. Note the values above each node or branch in the tree. These are called a bootstrap values, and are a measure of statistical confidence for each node. Any bootstrap score > 70 is typically considered as fairly reliable.
 - In the Tree Explorer window, go to **View > Options** (or click on the  icon) and choose the Branches option and then expand the Statistic/Frequency section.
 - Select the box that says **Hide values lower than** and input 70%
 - a. How many bootstrap reliable nodes are there?
 - b. Are more reliable nodes found nearer the base or the terminal tips of the tree?
 - c. Can you think of possible reasons for this?
 - d. Do you have much confidence in this tree?

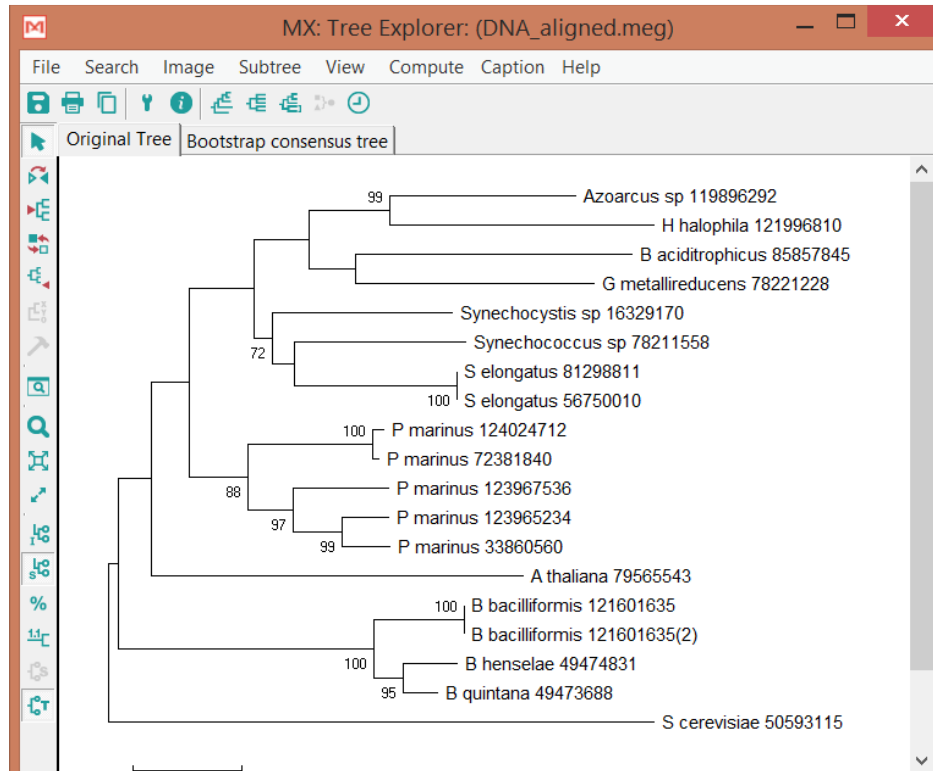


Figure 3. Bootstrapped Neighbour-Joining Tree displaying only those bootstrap scores $\geq 70\%$

5. You can change the way the way the tree is presented very simply in MEGA.
 - In the Tree Explorer window, go to **View > Tree Branch Style** and select some different formats such as circular, radial, traditional straight...
 - Examine the relative branching order of the sequences with these different formats
 - *Do the relationships of the sequences change?*
6. The MEGA Tree Explorer is very powerful. You can manipulate your tree endlessly through the **Options** (View > Options) and **Subtree** menus. Return your tree to the traditional / rectangular format and try playing with some of the control options. Most of the **Subtree** options are also available as icons on the left side of the window. Note that none of these changes are irreversible, so don't be shy about playing with them!

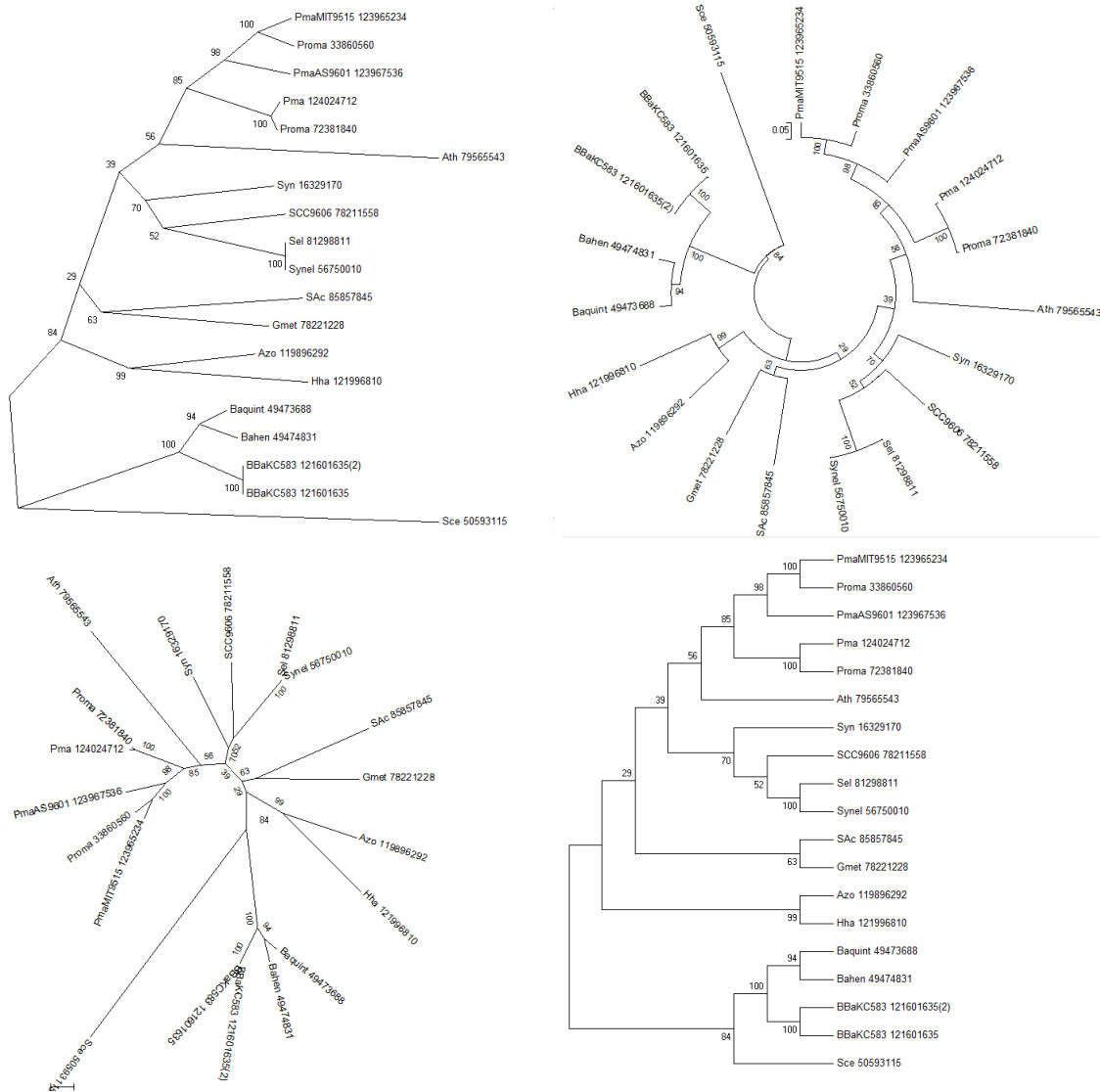


Figure 4. Four formats for the same tree. From top left to bottom right: traditional straight, circle, radial, topology only.

Box 4. Rooting Phylogenetic Trees

The *root* of a phylogenetic tree is exactly what it sounds like it should be – that point farthest back in time at the base of the tree. While this is a simple idea, determining the root of a tree can actually be very difficult. There are two primary ways to root trees:

1. Midpoint rooting involves placing the root at what is effectively the center of gravity of the tree. This is the default method used in MEGA and many other programs. Midpoint root is very easy to do (click **View > Root on Midpoint**), but

makes the very strong assumption that all of the sequences are evolving at approximately the same rate. If this is not the case then midpoint rooting can be inappropriate.

2. Outgroup rooting involves including a sequence that is known to be more divergent than the rest of the sequences in your analysis, and then making sure that it branches at the very base of the tree. Outgroup rooting is very reliable if you have solid prior information that permits you to choose a good outgroup. Unfortunately this information is not known for many studies.

If you are unsure about the rooting of the tree you can always present an *unrooted* tree, which draws the tree radiating out from a center point with no direction representing time (the radial tree shown above). Many people find these trees to be harder to interpret, but they present the same information, without the added assumption about which sequences branched first.

7. Let's see how changing the rooting of the tree affects our conclusions. Left click directly on one of the interior edges of the tree (that edge will be enclosed in a green box), and then right click to select **Place Root**. The tree will rearrange so that the edge you selected is now at the base of the tree.
 - Repeat a number of times on different branches and examine the relationships of some of the sequences.
 - a. *Do any of the relationships change?*
 - You can return to the midpoint rooting by selecting **View / Root on Midpoint**.

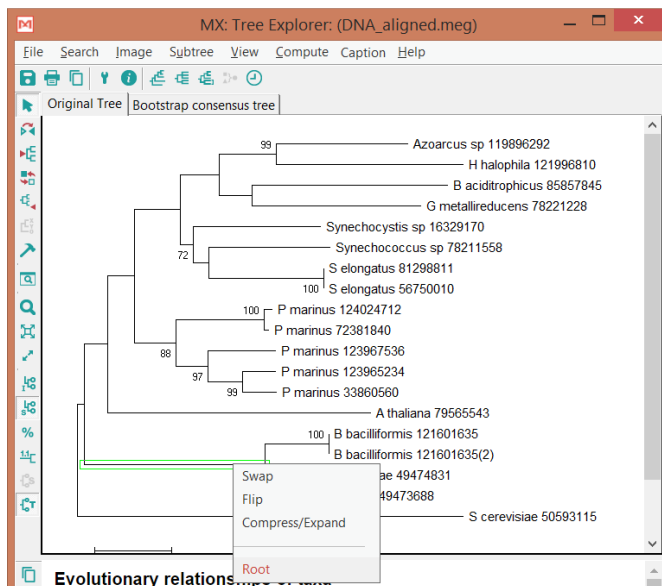


Figure 5. Re-rooting a tree in MEGA: click the branch to select it, then use the right mouse button to choose “Root”

8. Draw an unrooted tree by selecting **View > Tree Branch Style > Radiation**.
 - a. *Try to reconcile this tree with the one you were previously looking at.*

Lab Quiz
Question 1

9. MEGA X also produces an automatic caption for each analysis. Select **Caption** to see the detailed explanation of your particular phylogeny (the default behaviour is to display this below the tree). While you probably don't want to use this caption "as is" in a publication, this feature very clearly provides all the necessary information to describe your analysis.

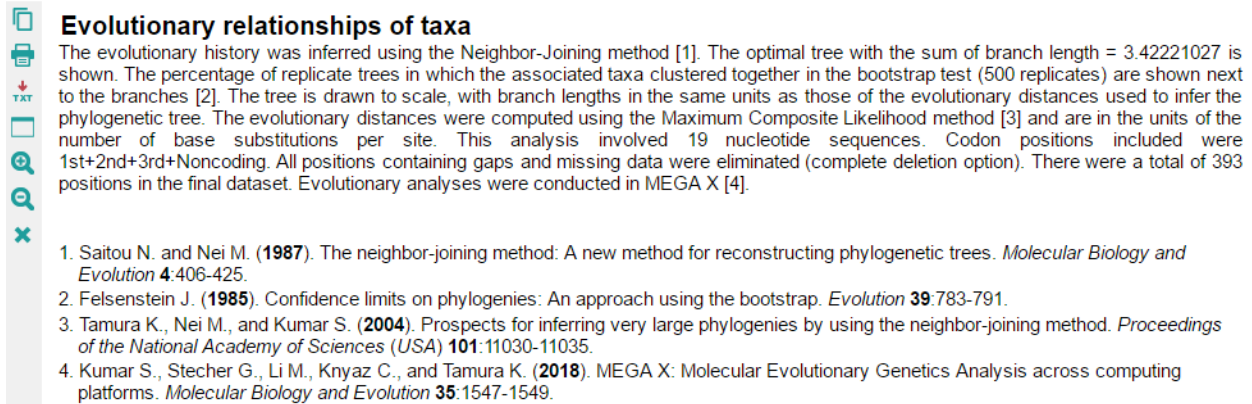


Figure 6. Auto-caption feature of MEGA X. Notice how bioinformatic software/methods are nicely cited!

10. Now let's make a tree with the protein sequences. You can keep the Tree Explorer window open, but close the current data file by clicking on the **Close Data** icon in the main window. Use the **corresponding aligned protein sequence file** (the "bioinformethods1-labs-Lab3,4_sequences_prot_aligned.fas" file from the Coursera section where you downloaded this lab from): you may have generated this alignment yourself last lab, but for consistency, use the provided alignment). You will need to convert your aligned protein sequences to MEGA format first. See notes under Step 1 on Page 3 for tips on doing this. Don't forget to open this file as per Step 2 on Page 4!
11. Make another neighbour-joining tree, the same way as in Step 3 (neighbour-joining defaults for protein alignments are in Appendix 2). Remove low bootstraps the same way you did previously. Compare the nucleotide and protein trees.
- a. Are there differences between these trees?
 - b. If so, why might this be?
 - c. Is one tree more "trustworthy" than the other?
12. Now let's play with some of the tree building parameters to see how they influence the analysis. Open up your aligned nucleotide sequence file again (note that you can do this by simply clicking on the **Close Data** icon and then selecting **File > Open a Recently Used File**), and then select **Phylogeny > Construct/Test Neighbor-Joining Tree...** (see Figure 2).
- In the parameters window, modify some of the parameters and rebuild the tree. Remember, you can't break anything, so try options out as much as possible. One very good way to determine if one tree is better than another is to look at the bootstrap scores. Try to optimize the parameters so that you have the strongest node

Lab Quiz
Question 2

support (higher bootstrap scores) for as many nodes as possible. Pay particular attention to the following:

- **Gaps-Missing Data / Pairwise Deletion**
 - This parameter is useful if you have sequences with lots of insertions and deletions (indels). Complete deletion (the default setting) removes all alignment columns with an indel in any sequence from the analysis. Pairwise deletion only removes alignment columns with indels from the analysis in pairwise comparisons. While complete deletion is the most conservative approach, sometimes the numbers of indels is so high that you lose significant information with this approach.
 - **Model / Nucleotide**
 - These substitution models are just like the substitution matrices discussed in previous labs. Try using different models to see if they influence your conclusions. See **Box 5** for a description of some of the models.
 - **Rates among sites / Different (Gamma Distributed)**
 - **Gamma Parameter** / try from 0.1 – 2.0
 - This controls for variation in the rate of evolution across a sequence. For example, maybe one region is highly conserved (evolves very slowly), while another is not conserved at all (has a much higher rate of evolution).
 - Lower gamma parameters are used for sequences with higher rate variation.
- a. *Describe how changing the parameters influence the phylogenetic reconstruction.*
 - b. *Can you make an educated guess why you might or might not want to include or exclude specific sites such as 3rd position in a codon or non-coding sites?*

Box 5. Substitution Models

As discussed previously, substitution models are used for modeling how DNA or protein sequences change over evolutionary time. While matrices like PAM and BLOSUM are useful for modeling protein sequence evolution, other models are used for DNA sequences. Here is a small selection of the models and their assumptions:

- Jukes-Cantor
 - equal frequencies for all nucleotides
 - no bias in the frequency in which any one mutates to another (equal substitution rates)
- Felsenstein-81
 - unequal frequencies for all nucleotides
 - equal substitution rates
- Kimura 2-Parameter
 - equal frequencies for all nucleotides
 - different substitution rates for transitions (purine to purine, or pyrimidine to pyrimidine) and transversions (purine to pyrimidine or vice versa). It is not uncommon for transitions to occur about twice as frequently as transversions.
- Tajima-Nei
 - unequal frequencies for all nucleotides
 - equal transversion frequencies
 - variable transition frequencies
- Tamura 3-Parameter
 - equal frequencies for all nucleotides
 - different substitution rates for transitions and transversions
 - G+C content bias
- Hasegawa-Kishino-Yano (HKY)
 - unequal frequencies for all nucleotides
 - different substitution rates for transitions and transversions

How do you choose which model to use? Most importantly, you need to look at your data. This should give you an idea if nucleotides are at different frequencies, and the appropriate transition – transversion rate. If you are really serious about doing it properly then use a program called *jModelTest 2*, which uses a likelihood approach to help you determine the best substitution model and gamma parameter. This program is beyond the scope of this course but may be accessed as a Java application at <https://code.google.com/p/jmodeltest2/> (Darriba et al., 2012).

Box 6. Character-Based Methods

Again, there are numerous character-based phylogenetic methods that work in very different manners. All of these approaches actually compare the state of each residue (nucleotide or amino acid) in each alignment column in a MSA. They attempt to identify the most likely or simplest explanation required to explain the relationships observed in the data. They generally do this by looking at all possible explanations (in other words, all possible trees), and identifying the single tree or set of trees that explains the data the best based on the specific criteria used in the method.

Maximum likelihood actually describes a statistical framework that in this case is applied to phylogenetic reconstruction. It basically goes through every possible tree structure and asks how likely your particular dataset is given a particular tree. So, for example, it would be much more likely that very similar sequences would branch very close together (near the tip of the tree), rather than branch very far apart (near the base of the tree).

Character-based methods such as maximum likelihood are generally very sophisticated approaches that permit realistic modeling of evolutionary change in a statistical framework. Unfortunately, they can also be more difficult to run (or at least run properly), and perhaps most importantly, they are computationally intensive since they effectively examine every possible tree structure. This practically means that they cannot be applied to very large datasets.

Now let's make a Maximum Likelihood (ML) tree. As discussed above, ML is one of the most powerful phylogenetic methods, but unfortunately, it is not quite as easy to perform as neighbour-joining. There are a number of good applications for ML analysis freely available. We will use the implementation in MEGA, although you might also want to familiarize yourself with the ML tools available through PHYLIP (Phylogeny Inference Package), which is a powerful and comprehensive set of freely available phylogenetic applications. PHYLIP runs on most computer platforms via a command interface, but is also available through a number of publically available web interfaces (see "Where to get it", below).

1. Start MEGA and load the aligned DNA sequence in MEGA format as per Step 2 on Page 4 (**File > Open a Recently Used File** and choose the aligned DNA sequences in MEGA format that you prepared earlier in this lab).
2. Under **Analysis > Phylogeny > Construct/Test Maximum Likelihood Tree...** set the **Phylogeny Test** / Test of Phylogeny to Bootstrap, with 500 replications, and leave all the other settings at their default values (see Appendix 2 for defaults; increasing the number of threads might speed up the analysis, however). Click **Compute** to start the analysis.
3. Wait quite a bit longer than you did with the Neighbour Joining analysis. As explained in the mini-lecture, a lot more computation is used in the Maximum Likelihood algorithm. The progress bar will show the extent to which your analysis is complete. Once the analysis is done, the Tree Viewer will open and display the resultant tree.

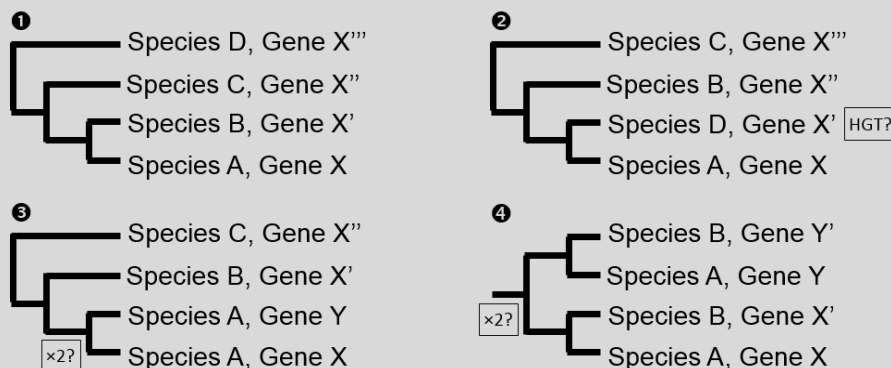
4. You can now manipulate and play with the tree just as you did previously.
 - a. *What are the differences between this tree and others that you have created?*
 - b. *Compare the branching order of the sequences from this tree and the others. Do they differ?*

Lab Quiz
Question 3

- One of the strongest ways of supporting a phylogenetic analysis is to perform it using at least two independent methods. If you obtain the same basic topology using both neighbour-joining and ML, for example, then you have very good reason to believe that your analysis is sound. Nearly every molecular evolution journal requires that all phylogenetic analyses be performed using multiple approaches.
5. If you have time, go back and play with some of the Maximum Likelihood analysis options in MEGA to see how they influence the tree structure.

Box 7. Interpreting Phylogenetic Analyses

Phylogenetic analyses are very powerful for determining the evolutionary history of a gene of interest. Consider the following four scenarios, and assume that the nodes all have good (high levels of) bootstrap support. Genes X, X', X'', X''' are orthologs of each other, and Gene Y and Y' are paralogous to these genes. Assume Species A, B, C and D are more distantly related to each other the further along the alphabet one goes. In Scenario ①, the gene falls within its expected clade, that is, the gene tree mirrors the species tree. No surprises here! In Scenario ②, the gene doesn't fall where we expect it relative to the species tree. This gene may have been acquired by a horizontal transfer event [HGT?] from the species (or a closely related species) with which the gene is grouping. In Scenario ③, there is a paralog in the species that the gene is from, but there are no paralogs in other species. Assuming that the genomes of these other species have been sequenced and our E-value threshold wasn't too stringent, this may be indicative of a partial or whole genome duplication event having occurred in the species for the gene we're interested in, denoted by [x2?]. In Scenario ④, the paralogs of the gene also have homologs in other species. Again, with the caveat of good genome coverage and the proper E-value cutoff, this may indicate that the duplication event happened in an ancestor of both species, at the point denoted by [x2?].



End of Lab!

Where to get it:

MEGA X	http://www.megasoftware.net/
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
Web-based ML analysis	http://bar.utoronto.ca/webphylip/

Lab 4 Objectives

By the end of Lab 4 (comprising the lab including its boxes, and the lecture), you should:

- know the terminology associated with phylogenetic trees and be able to identify the most recent common ancestor of any two terminal nodes (taxa) on the tree;
- know the fundamental elements and the terminology of phylogenetics, and possible evolutionary routes to a given derived state (how can homoplasy arise?);
- be able to identify the root of a tree and to know the difference between rooted and unrooted trees;
- understand distance and character-based phylogenetic methods, how they work, and the pros and cons of each;
- be acquainted with the various substitution models;
- be familiar with bootstrapping and what a bootstrap score on a node tells you;
- be able practically to construct neighbour-joining and maximum likelihood trees using MEGA.

Do not hesitate to check with the Forums for this course on Coursera if you do not understand any of the above after reading the relevant material.

Further Reading

Chapters 8 “Phylogenetics” in *Concepts in Bioinformatics and Genomics* by Jamil Momand and Alison McCurdy, Oxford University Press, 2017. pp. 168-209.

Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

WF Doolittle (1999) Phylogenetic classification and the universal tree. *Science* 284: 2124-2128.

RDM Page and MA Charleston (1997) From gene to organismal phylogeny: reconcile trees and the gene / species tree problem. *Mol. Phylogenet. Evol.* 7:231-240.

N Saitou and M Nei (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406-425.

S Guidon and O Gascuel (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52: 696-704.

T Sitnikova (1996) Bootstrap method of interior-branch test for phylogenetic trees. *Mol. Biol. Evol.* 13: 605-611.

Appendix 1: File Formats

Fasta Format

```
>A_thaliana_79565543
ATCAGCGATATCCCAAGAAGAACAAAGTTTCAGAAACATCATCGAGGAAGAATTAATAAA
GGAGTATCTTCTCAGGGGTATATTTGTAGTAGATATGCTCTTCAAACACTTGAACCAGCT
TGGATCACTTCTAGACAAATAGAAGCAGGACGACGAGCAATGAC
>Azoarcus_sp_119896292
ATGCTGCAGCCGTCGAGAAGGAAATACCGCAAGGAGCAGAAAGGTCGCAACACCGGCCTG
GCGACGCGCGGCACCAAGGTCAGCTTCGGTGATTTTCGGTCTGAAGGCGATCGCCCGCGGT
CGTCTGACTGCCCCGTCAGATTGAATCCGCGCGTCGCGCGATGAC
>P_marinus_124024712
ATGCTTAGCCCAAAAAGAACCAAATTTTCGTAAACAACAAAGAGGCCGTATGCGCGGTGTT
GCTACTAGAGGCAACAAAATCGCTTTTGGTCAGTTTGCATTGCAAGCTCAAGACTGTGGA
TGGGTCACCTCAAGGCAAATCGAGGCAAGTCGACGAGCAATGAC
>H_halophila_121996810
ATGTTACAGCCGAAACGGACCAAGTACCGCAAGAAGCAAAGGGCCGCTGCTCGGGCCTC
GCGACCCGCGGTGATCGCGTGAGCTTCGGCGAGTTTCGGCCTCAAGGCAACCACCCGCGGG
CCGATCACCTCGCGGCAGATCGAGGCGGCGCGGCGTGCCATCAA
>B_bacilliformis_121601635
ATGTTGCAGCCAAAGCGCACAAAGTTCGGTAAGCAATTCAAAGGTCGTATTCACGGTGCT
TCGAAAGGTGGTACGGATTTGAATTTTGGTGCTTACGGCCTGAAGGTTGTCGAGCCAGAG
CGTATTACTGCGCGTCAAATTGAAGCAGCTCGTCGTGCAATTAC
>S_aciditrophicus_85857845
ATGTTAATGCCAAAAGGGTGAAATATAGGAAGTTGCAAAGGGGTGCAAGGACAGGAACC
GCCACAAGAGGAAGTAAAATATCTTTTGGGGAATATGGACTTCAAGCAGAAGAATGTGGC
TGGATAACCGCAAGGCAGATTGAGGCAGCGAGAATTGCCATTAC
>S_elongatus_81298811
ATGCTCAGTCCACGTCTGTACCAAATTCGGGAAGCAGCAACGTGGCCGCATGACCGGCAAA
GCGACGCGCGGGAATACTCTCGCCTTCGGTAACTTCGGTCTGCAGGCGCTGGAATGCTCC
TGGATCACGGCTCGCCAAATTGAGGCTAGCCGTCGTGCCATGAC
>G_metallireducens_78221228
ATGTTGATGCCAAAAGAGTTAAGTATAGAAAGCAAATGAAGGGGCGCATGACGGGCGCT
GCAATGCGCGGGGCCACACTGTCTGACGGTGATTTTCGGTCTCCAGGCAACGGAGTGTTGA
TGGGTTGATTTCCCGTCAGATAGAGGCTGCTCGTATTGCAATGAC
>Synechococcus_sp_78211558
ATGCTGAGTCCAAAACGCGTCAAATTCGGTAAGCAGCAGCGAGGCCGCATGCGCGGCGTC
GCCACCCGGGGCAACACCATTGCCTTCGGACAATTTCGCGCTGCAGGCACAGGAATGTGGC
TGGATCACCTCGCGCCAGATCGAGGCCAGCCGTCGTGCCATGAC
>B_henselae_49474831
ATGTTGCAGCCAAAGCGCACAAAGGTTCCGTAAACAGTTCAAAGGTCGTATTCATGGTGTT
TCGAAAGGTGGTACGGATTTGAATTTTCGGTGCTTATGGTTTAAAAGCTGTTGAACCGGAG
CGGATTACTGCCCCGCAAATTGAGGCGGCGCGTCGTGCGATTAC
>B_quintana_49473688
ATGTTGCAGCCAAAGCGCACAAAGGTTCCGTAAACAATTCAAAGGTCGTATTCACGGTGTT
TCGAAAGGTGGTACGGATCTAAATTTTCGGTGCTTATGGTTTAAAAGCTGTTGAGCCGGAG
CGGATTACTGCCCCGCAAATTGAAGCGGCGCGTCGTGCGATTAC
```

MEGA Format

```

#mega
TITLE: Written by EMBOSS 27/01/09
#A_thaliana_79565543 ATCAGCGATATCCCAAGAAGAACAAAGTTTCAGAAACATCATCGAGGAAG
#Azoarcus_sp_11989629 ATGCTGCAGCCGTCGAGAAGGAAATACCGCAAGGAGCAGAAAGGTCGCAA
#P_marinus_124024712 ATGCTTAGCCCAAAAAGAACCAATTTTCGTAAACAACAAAGAGGCCGTAT
#H_halophila_12199681 ATGTTACAGCCGAAACGGACCAAGTACCGCAAGAAGCAAAAGGGCCGCTG
#B_bacilliformis_1216 ATGTTGCAGCCAAAGCGCACAAAGTTCCGTAAGCAATTCAAAGGTCGTAT
#S_aciditrophicus_858 ATGTTAATGCCAAAAGGGTGAAATATAGGAAGTTGCAAAGGGGTGCAAG
#S_elongatus_81298811 ATGCTCAGTCCACGTCGTACCAAATTCGGAAGCAGCAACGTGGCCGCAT
#G_metallireducens_78 ATGTTGATGCCAAAAGAGTTAAGTATAGAAAGCAAATGAAGGGGCGCAT
#Synechococcus_sp_782 ATGCTGAGTCCAAAACGCGTCAAATTCGTAAGCAGCAGCGAGGCCGCAT
#B_henselae_49474831 ATGTTGCAGCCAAAGCGCACAAAGGTTCCGTAAACAGTTCAAAGGTCGTAT
#B_quintana_49473688 ATGTTGCAGCCAAAGCGCACAAAGGTTCCGTAAACAAATTCAAAGGTCGTAT
#A_thaliana_79565543 AATTAATAAAGAGTATCTTCTCAGGGGTATATTTGTAGTAGATATGCTC
#Azoarcus_sp_11989629 CACCGGCCTGGCGACGCGCGGCCACCAAGGTCAGCTTCGGTGATTTTCGGTC
#P_marinus_124024712 GCGCGGTGTTGCTACTAGAGGCAACAAAATCGCTTTTGGTCAGTTTGCAT
#H_halophila_12199681 CTCGGGCCTCGCGACCCGCGGTGATCGCGTGAGCTTCGGCGAGTTTCGGCC
#B_bacilliformis_1216 TCACGGTGCTTCGAAAGGTGGTACGGATTTGAATTTTGGTGCTTACGGCC
#S_aciditrophicus_858 GACAGGAACCGCCACAAGAGGAAGTAAAATATCTTTTGGGGAATATGGAC
#S_elongatus_81298811 GACCGGCAAAGCGACGCGCGGGAATACTCTCGCCTTCGGTAACCTTCGGTC
#G_metallireducens_78 GACGGGCGCTGCAATGCGCGGGGCCACACTGTCTGACGGTGATTTTCGGTC
#Synechococcus_sp_782 GCGCGGCGTCGCCACCCGGGGCAACACCATTGCCTTCGGACAATTTCGCGC
#B_henselae_49474831 TCATGGTGTTTCGAAAGGTGGTACGGATTTGAATTTTCGGTGCTTATGGTT
#B_quintana_49473688 TCACGGTGTTTCGAAAGGTGGTACGGATCTAAATTTTCGGTGCTTATGGTT
#A_thaliana_79565543 TTCAAACACTTGAACCAGCTTGGATCACTTCTAGACAAATAGAAGCAGGA
#Azoarcus_sp_11989629 TGAAGGCGATCGCCCGCGGTCTGCTGACTGCCCCTCAGATTGAATCCGCG
#P_marinus_124024712 TGCAAGCTCAAGACTGTGGATGGGTCACTTCAAGGCAAATCGAGGCAAGT
#H_halophila_12199681 TCAAGGCAACCACCCGCGGGCCGATCACCTCGCGGCAGATCGAGGCGGCG
#B_bacilliformis_1216 TGAAGGTTGTGAGCCAGAGCGTATTACTGCGCGTCAAATTGAAGCAGCT
#S_aciditrophicus_858 TTCAAGCAGAAGAATGTGGCTGGATAACCGCAAGGCAGATTGAGGCAGCG
#S_elongatus_81298811 TGCAGGCGCTGGAATGCTCCTGGATCACGGCTCGCCAAATTGAGGCTAGC
#G_metallireducens_78 TCCAGGCAACGGAGTGTGGATGGGTTGATTCCCGTCAGATAGAGGCTGCT
#Synechococcus_sp_782 TGCAGGCACAGGAATGTGGCTGGATCACCTCGCGCCAGATCGAGGCCAGC
#B_henselae_49474831 TAAAAGCTGTTGAACCGGAGCGGATTACTGCCCGCCAAATTGAGGCGGCG
#B_quintana_49473688 TGAAAGCTGTTGAGCCGGAGCGGATTACTGCCCGCCAAATTGAAGCGGCG
#A_thaliana_79565543 CGACGAGCAATGAC
#Azoarcus_sp_11989629 CGTCGCGCGATGAC
#P_marinus_124024712 CGACGAGCAATGAC
#H_halophila_12199681 CGGCGTGCCATCAA
#B_bacilliformis_1216 CGTCGTGCAATTAC
#S_aciditrophicus_858 AGAATTGCCATTAC
#S_elongatus_81298811 CGTCGTGCCATGAC
#G_metallireducens_78 CGTATTGCAATGAC
#Synechococcus_sp_782 CGTCGTGCCATGAC
#B_henselae_49474831 CGTCGTGCGATTAC
#B_quintana_49473688 CGTCGTGCGATTAC

```


Clustal Format

```

A_thaliana_7956 ATCAGCGATATCCCAAGAAGAACAAAGTTTCAGAAACATCATCGAGGAAGAATTAATAAA
Azoarcus_sp_119 ATGCTGCAGCCGTCGAGAAGGAAATACCGCAAGGAGCAGAAAGGTCGCAACACCGGCCTG
P_marinus_12402 ATGCTTAGCCCCAAAAGAACCAAAATTTTCGTAAACAACAAAGAGGCCGTATGCGCGGTGTT
H_halophila_121 ATGTTACAGCCGAAACGGACCAAGTACCGCAAGAAGCAAAAGGGCCGCTGCTCGGGCCTC
B_bacilliformis ATGTTGCAGCCAAAGCGCACAAAGTTCGGTAAGCAATTCAAAGGTCGTATTCACGGTGCT
S_aciditrophicu ATGTTAATGCCAAAAGGGTGAAATATAGGAAGTTGCAAAGGGGTGCAAGGACAGGAACC
S_elongatus_812 ATGCTCAGTCCACGTTCGTACCAAATTCGGGAAGCAGCAACGTGGCCGCATGACCGGCAAA
G_metallireduce ATGTTGATGCCAAAAGAGTTAAGTATAGAAAGCAAATGAAGGGGCGCATGACGGGCGCT
Synechococcus_s ATGCTGAGTCCAAAACGCGTCAAATTCGGTAAGCAGCAGCGAGGCCGCATGCGCGGCGTC
B_henselae_4947 ATGTTGCAGCCAAAGCGCACAAAGTTCGGTAACAGTTCAAAGGTCGTATTCATGGTGTT
B_quintana_4947 ATGTTGCAGCCAAAGCGCACAAAGTTCGGTAACAAATTCAAAGGTCGTATTCACGGTGTT

A_thaliana_7956 GGAGTATCTTCTCAGGGGTATATTTGTAGTAGATATGCTCTTCAAACACTTGAACCAGCT
Azoarcus_sp_119 GCGACGCGCGGCACCAAGGTCAGCTTCGGTGATTTTCGGTCTGAAGGCGATCGCCCGCGGT
P_marinus_12402 GCTACTAGAGGCAACAAAATCGCTTTTGGTCAGTTTGCATTGCAAGCTCAAGACTGTGGA
H_halophila_121 GCGACCCGCGGTGATCGCGTGAGCTTCGGCGAGTTTCGGCCTCAAGGCAACCACCCGCGGG
B_bacilliformis TCGAAAGGTGGTACGGATTTGAATTTTGGTGCTTACGGCCTGAAGGTTGTGAGCCAGAG
S_aciditrophicu GCCACAAGAGGAAGTAAAATATCTTTTGGGGAATATGGACTTCAAGCAGAAGAATGTGGC
S_elongatus_812 GCGACGCGCGGGAATACTCTCGCCTTCGGTAACCTTCGGTCTGCAGGCGCTGGAATGCTCC
G_metallireduce GCAATGCGCGGGGCCACACTGTCTGACGGTGATTTTCGGTCTCCAGGCAACGGAGTGTTGA
Synechococcus_s GCCACCCGGGGCAACACCATTGCCTTCGGACAATTTCGCGCTGCAGGCACAGGAATGTGGC
B_henselae_4947 TCGAAAGGTGGTACGGATTTGAATTTTCGGTGCTTATGGTTTAAAAGCTGTTGAACCGGAG
B_quintana_4947 TCGAAAGGTGGTACGGATCTAAATTTTCGGTGCTTATGGTTTGAAGCTGTTGAGCCGGAG

A_thaliana_7956 TGGATCACTTCTAGACAAATAGAAGCAGGACGACGAGCAATGAC
Azoarcus_sp_119 CGTCTGACTGCCCCGTCAGATTGAATCCGCGCGTCGCGCGATGAC
P_marinus_12402 TGGGTCACTTCAAGGCAAATCGAGGCAAGTCGACGAGCAATGAC
H_halophila_121 CCGATCACCTCGCGGCAGATCGAGGCGGCGCGCGTGCATCAA
B_bacilliformis CGTATTACTGCGCGTCAAATTGAAGCAGCTCGTCGTGCAATTAC
S_aciditrophicu TGGATAACCGCAAGGCAGATTGAGGCAGCGAGAATTGCCATTAC
S_elongatus_812 TGGATCACGGCTCGCCAAATTGAGGCTAGCCGTCGTGCCATGAC
G_metallireduce TGGGTTGATTCCCGTCAGATAGAGGCTGCTCGTATTGCAATGAC
Synechococcus_s TGGATCACCTCGCGCCAGATCGAGGCCAGCCGTCGTGCCATGAC
B_henselae_4947 CGGATTACTGCCCCGCCAAATTGAGGCGGCGCGTCGTGCGATTAC
B_quintana_4947 CGGATTACTGCCCCGCCAAATTGAAGCGGCGCGTCGTGCGATTAC

```

PHYLIP Interleaved Format

11 164

```

A_thalianaATCAGCGATA TCCCAAGAAG AACAAAGTTT CAGAAACATC ATCGAGGAAG
Azoarcus_sATGCTGCAGC CGTCGAGAAG GAAATACCGC AAGGAGCAGA AAGGTCGCAA
P_marinus_ATGCTTAGCC CAAAAAGAAC CAAATTTTCGT AAACAACAAA GAGGCCGTAT
H_halophilATGTTACAGC CGAAACGGAC CAAGTACCGC AAGAAGCAAA AGGGCCGCTG
B_bacillifATGTTGCAGC CAAAGCGCAC AAAGTTCCGT AAGCAATTCA AAGGTCGTAT
S_aciditroATGTTAATGC CAAAAAGGGT GAAATATAGG AAGTTGCAA GGGGTGCAAG
S_elongatuATGCTCAGTC CACGTCGTAC CAAATTCCGG AAGCAGCAAC GTGGCCGCAT
G_metallirATGTTGATGC CAAAAGAGT TAAGTATAGA AAGCAAATGA AGGGGCGCAT
SynechococATGCTGAGTC CAAAACGCGT CAAATTCCGT AAGCAGCAGC GAGGCCGCAT
B_henselaeATGTTGCAGC CAAAGCGCAC AAGGTTCCGT AAACAGTTCA AAGGTCGTAT
B_quintanaATGTTGCAGC CAAAGCGCAC AAGGTTCCGT AAACAATTCA AAGGTCGTAT

```

```

AATTAATAAA GGAGTATCTT CTCAGGGGTA TATTTGTAGT AGATATGCTC
CACCGGCCTG GCGACGCGCG GCACCAAGGT CAGCTTCGGT GATTTTCGGT
GCGCGGTGTT GCTACTAGAG GCAACAAAAT CGCTTTTGGT CAGTTTGCAT
CTCGGGCCTC GCGACCCGCG GTGATCGCGT GAGCTTCGGC GAGTTCGGCC
TCACGGTGCT TCGAAAGGTG GTACGGATTT GAATTTTGGT GCTTACGGCC
GACAGGAACC GCCACAAGAG GAAGTAAAAT ATCTTTTGGG GAATATGGAC
GACCGGCAAA GCGACGCGCG GGAATACTCT CGCCTTCGGT AACTTCGGTC
GACGGGCGCT GCAATGCGCG GGGCCACACT GTCGTACGGT GATTTTCGGT
GCGCGGCGTC GCCACCCGGG GCAACACCAT TGCCTTCGGA CAATTTCGCGC
TCATGGTGTT TCGAAAGGTG GTACGGATTT GAATTTTCGGT GCTTATGGTT
TCACGGTGTT TCGAAAGGTG GTACGGATCT AAATTTTCGGT GCTTATGGTT

```

```

TTCAAACACT TGAACCAGCT TGGATCACTT CTAGACAAAT AGAAGCAGGA
TGAAGGCGAT CGCCCGCGGT CGTCTGACTG CCCGTCAGAT TGAATCCGCG
TGCAAGCTCA AGACTGTGGA TGGGTCACTT CAAGGCAAAT CGAGGCAAGT
TCAAGGCAAC CACCCGCGGG CCGATCACCT CGCGGCAGAT CGAGGCGGCG
TGAAGGTTGT CGAGCCAGAG CGTATTACTG CGCGTCAAAT TGAAGCAGCT
TTCAAGCAGA AGAATGTGGC TGGATAACCG CAAGGCAGAT TGAGGCAGCG
TGCAGGCGCT GGAATGCTCC TGGATCACGG CTCGCCAAAT TGAGGCTAGC
TCCAGGCAAC GGAGTGTGGA TGGGTTGATT CCCGTCAGAT AGAGGCTGCT
TGCAGGCACA GGAATGTGGC TGGATCACCT CGCGCCAGAT CGAGGCCAGC
TAAAAGCTGT TGAACCGGAG CGGATTACTG CCCGCCAAAT TGAGGCGGCG
TGAAAGCTGT TGAGCCGGAG CGGATTACTG CCCGCCAAAT TGAAGCGGCG

```

```

CGACGAGCAA TGAC
CGTCGCGCGA TGAC
CGACGAGCAA TGAC
CGGCGTGCCA TCAA
CGTCGTGCAA TTAC
AGAATTGCCA TTAC
CGTCGTGCCA TGAC
CGTATTGCAA TGAC
CGTCGTGCCA TGAC
CGTCGTGCGA TTAC
CGTCGTGCGA TTAC

```

PHYLIP Non-interleaved Format

11 164

```

A_thalianaATCAGCGATA TCCCAAGAAG AACAAAGTTT CAGAAACATC ATCGAGGAAG
      AATTAATAAA GGAGTATCTT CTCAGGGGTA TATTTGTAGT AGATATGCTC
      TTCAAACACT TGAACCAGCT TGGATCACTT CTAGACAAAT AGAAGCAGGA
      CGACGAGCAA TGAC
Azoarcus_sATGCTGCAGC CGTCGAGAAG GAAATACCGC AAGGAGCAGA AAGGTCGCAA
      CACCGGCCTG GCGACGCGCG GCACCAAGGT CAGCTTCGGT GATTTTCGGTC
      TGAAGGCGAT CGCCCGCGGT CGTCTGACTG CCCGTCAGAT TGAATCCGCG
      CGTCGCGCGA TGAC
P_marinus_ATGCTTAGCC CAAAAAGAAC CAAATTTTCGT AAACAACAAA GAGGCCGTAT
      GCGCGGTGTT GCTACTAGAG GCAACAAAAT CGCTTTTGGT CAGTTTGCAT
      TGCAAGCTCA AGACTGTGA TGGGTCCTT CAAGGCAAAT CGAGGCAAGT
      CGACGAGCAA TGAC
H_halophilATGTTACAGC CGAAACGGAC CAAGTACCGC AAGAAGCAAA AGGGCCGCTG
      CTCGGGCCTC GCGACCCGCG GTGATCGCGT GAGCTTCGGC GAGTTCGGCC
      TCAAGGCAAC CACCCGCGGG CCGATCACCT CGCGGCAGAT CGAGGCGGCG
      CGGCGTGCCA TCAA
B_bacillifATGTTGCAGC CAAAGCGCAC AAAGTTCCGT AAGCAATTCA AAGGTCGTAT
      TCACGGTGCT TCGAAAGGTG GTACGGATTT GAATTTTGGT GCTTACGGCC
      TGAAGGTTGT CGAGCCAGAG CGTATTACTG CGCGTCAAAT TGAAGCAGCT
      CGTCGTGCAA TTAC
S_aciditroATGTTAATGC CAAAAGGGT GAAATATAGG AAGTTGCAAA GGGGTCTGAAG
      GACAGGAACC GCCACAAGAG GAAGTAAAAT ATCTTTTGGG GAATATGGAC
      TTCAAGCAGA AGAATGTGGC TGGATAACCG CAAGGCAGAT TGAGGCAGCG
      AGAATTGCCA TTAC
S_elongatuATGCTCAGTC CACGTCGTAC CAAATTCGGG AAGCAGCAAC GTGGCCGCAT
      GACCGGCAAA GCGACGCGCG GGAATACTCT CGCCTTCGGT AACTTCGGTC
      TGCAGGCGCT GGAATGCTCC TGGATCACGG CTCGCCAAAT TGAGGCTAGC
      CGTCGTGCCA TGAC
G_metallirATGTTGATGC CAAAAGAGT TAAGTATAGA AAGCAAATGA AGGGGCGCAT
      GACGGGCGCT GCAATGCGCG GGGCCACACT GTCGTACGGT GATTTTCGGTC
      TCCAGGCAAC GGAGTGTGA TGGGTTGATT CCCGTCAGAT AGAGGCTGCT
      CGTATTGCAA TGAC
SynechococATGCTGAGTC CAAAACGCGT CAAATTCGGT AAGCAGCAGC GAGGCCGCAT
      GCGCGGCGTC GCCACCCGGG GCAACACCAT TGCCTTCGGA CAATTTCGCGC
      TGCAGGCACA GGAATGTGGC TGGATCACCT CGCGCCAGAT CGAGGCCAGC
      CGTCGTGCCA TGAC
B_henselaeATGTTGCAGC CAAAGCGCAC AAGGTTCCGT AAACAGTTCA AAGGTCGTAT
      TCATGGTGTT TCGAAAGGTG GTACGGATTT GAATTTTCGGT GCTTATGGTT
      TAAAAGCTGT TGAACCGGAG CGGATTACTG CCCGCCAAAT TGAGGCGGCG
      CGTCGTGCGA TTAC
B_quintanaATGTTGCAGC CAAAGCGCAC AAGGTTCCGT AAACAATTCA AAGGTCGTAT
      TCACGGTGTT TCGAAAGGTG GTACGGATCT AAATTTTCGGT GCTTATGGTT
      TGAAAGCTGT TGAGCCGGAG CGGATTACTG CCCGCCAAAT TGAAGCGGCG
      CGTCGTGCGA TTAC

```

Appendix 2: MEGA defaults for phylogenetic analyses for Lab 4

MEGA doesn't have a "reset" button to return to default parameters. The defaults for a nucleotide-based neighbour-joining analysis are shown in **Figure 2**. Others are below.

MX: Analysis Preferences

Option	Setting
ANALYSIS	
Scope	→ All Selected Taxa
Statistical Method	→ Neighbor-joining
PHYLOGENY TEST	
Test of Phylogeny	→ Bootstrap method
No. of Bootstrap Replications	→ 500
SUBSTITUTION MODEL	
Substitutions Type	→ Amino acid
Model/Method	→ Poisson model
RATES AND PATTERNS	
Rates among Sites	→ Uniform Rates
Gamma Parameter	→ Not Applicable
Pattern among Lineages	→ Same (Homogeneous)
DATA SUBSET TO USE	
Gaps/Missing Data Treatment	→ Complete deletion
Site Coverage Cutoff (%)	→ Not Applicable
SYSTEM RESOURCE USAGE	
Number of Threads	→ 3

? Help X Cancel ✓ OK

Neighbour-joining defaults for protein sequence alignments

MX: Analysis Preferences

Option	Setting
ANALYSIS	
Statistical Method	→ Maximum Likelihood
PHYLOGENY TEST	
Test of Phylogeny	→ Bootstrap method
No. of Bootstrap Replications	→ 500
SUBSTITUTION MODEL	
Substitutions Type	→ Nucleotide
Genetic Code Table	→ Not Applicable
Model/Method	→ Tamura-Nei model
RATES AND PATTERNS	
Rates among Sites	→ Uniform Rates
No of Discrete Gamma Categories	→ Not Applicable
DATA SUBSET TO USE	
Gaps/Missing Data Treatment	→ Complete deletion
Site Coverage Cutoff (%)	→ Not Applicable
Select Codon Positions	→ <input checked="" type="checkbox"/> 1st <input checked="" type="checkbox"/> 2nd <input checked="" type="checkbox"/> 3rd <input checked="" type="checkbox"/> Noncoding Sites
TREE INFERENCE OPTIONS	
ML Heuristic Method	→ Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	→ Make initial tree automatically (Default - NJ/BioNJ)
Initial Tree File	→ Not Applicable
Branch Swap Filter	→ Very Strong
SYSTEM RESOURCE USAGE	
Number of Threads	→ 3

? Help X Cancel ✓ OK

Maximum Likelihood defaults for nucleotide sequence alignments