Using the Free Program MEGA to Build Phylogenetic Trees from Molecular Data

(ABT-2015-0089) in the September issue of ABT (78,7), 2016

Lucas Newman

Amanda L. J. Duffus

Cathy Lee*

Department of Biology, Gordon State College, Barnesville, GA

* Corresponding Author: clee@gordonstate.edu

Abstract:

Building evolutionary trees can be an excellent way for students to see how different gene sequences or organisms are related to one another. Molecular Evolutionary Genetics Analysis (MEGA) software is a free package that lets anyone build evolutionary trees in a user-friendly setup. There are several different options to choose from when building trees from molecular data in MEGA, but the most commonly used are Neighbor-Joining and Maximum – Likelihood, both of which give good estimates on the relationship between different molecular sequences. In this article, we describe how to collect data from GenBank, insert it into a text editor, import that data into MEGA, and create phylogenetic trees from the collected data.

Key Words: MEGA, evolutionary trees, molecular data, Neighbor-Joining, Maximum-likelihood

Phylogenetics is the study of the evolutionary relatedness between different groups of organisms (Nei and Kumar 2000). These groups can be on small scales (e.g., mammals) or large

scales (e.g., different domains of life). The results of phylogenetic analyses are usually presented in the form of evolutionary trees, where different branches represent different gene sequences or species used to build them. The branching pattern of the tree illustrates how the sequences or species are related.

Here we present how to build evolutionary trees using the MEGA software package (newest version 7; www.megasoftware.net). It contains two of the most commonly used methods to infer evolutionary relationships among species by using their gene sequences. In this paper, we will introduce students to these two methods: Neighbor-Joining (NJ) and Maximum-Likelihood (ML) method. According to MEGA authors, it is frequently used in educational setting in advanced classes (pers. comm. Sudhir Kumar; Ryan et al. 2013). However, many K-12 instructors are not familiar with it, and its potential to introduce the concepts of evolutionary biology to students in a hands-on, discovery-based pedagogy using the gene sequences. There are multiple online resources that provide such gene sequences for a multitude of species (e.g., GenBank), which is available from National Center for Biotechnology Information (NCBI) (Hall 2013). Both DNA and protein sequences are available, and there are several informative tutorials provided on the NCBI website on how to use these. There is literally unlimited sequence data from thousands of genes from animals, plants, protists, bacteria, and viruses available through GenBank.

Project Goal:

To build an evolutionary tree using the rcbL gene sequence, which is commonly used to study the evolutionary relationships between plants (see Newmaster et al. 2006). Sequence data pertaining to the rcbL genes from many plant species are available through GenBank and we

will gather our data set from this resource. The rcbL gene sequence data will then be imported into MEGA, aligned, and then used to build a phylogenetic tree.

Helpful Prior Knowledge and Potential Context of this Exercise:

Students should have some introductory level knowledge of the purpose of evolutionary trees and have some experience interpreting simple phylogenetic trees. This exercise would be ideal as a final project for evolution units at varying levels, including AP Biology.

Learning Objectives:

By the end of this project, students will:

1. Learn how to obtain molecular data from GenBank (http://www.ncbi.nlm.nih.gov/genbank/).
2. Learn to build evolutionary trees using freely available software MEGA.
3. Understand the meaning of ancestral vs. recent species, clades, and be able to interpret evolutionary relationship among species.

As students perform the exercise, they should consider the following questions:

1. Why was the rcbL gene used?
2. Which organelle does the rbcL gene originate from?
3. What function does the protein product of the rcbL gene have in the plant?

System Requirements:

1. Internet access to use the GenBank data base and to download MEGA.
2. A text editor program, Notepad (Windows PC) or Texteditor (Linux/Mac). (http://www.megasoftware.net/mega.php)

Getting Started:

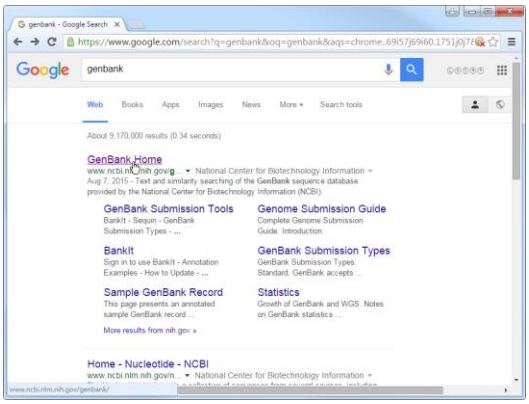


Figure 1. Google search for GenBank and click on the result for GenBank Home (http://www.ncbi.nlm.nih.gov/genbank/).

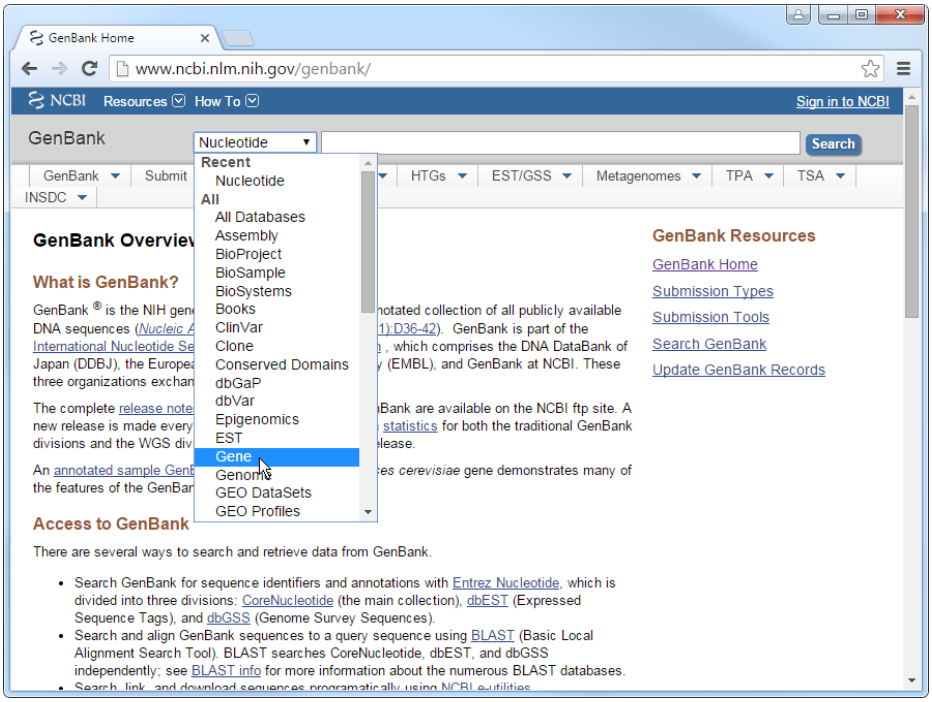


Figure 2. Use the dropdown next to the word GenBank to change from default Nucleotide and select Gene.

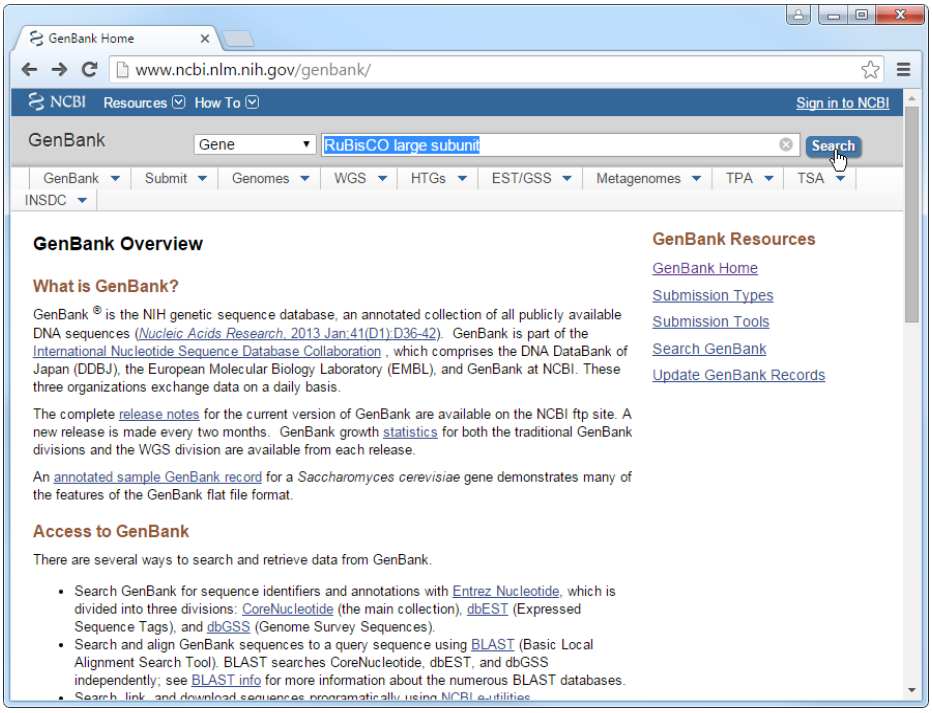


Figure 3. Type **RuBisCO large subunit** in the entry box to the right of dropdown and click Search.

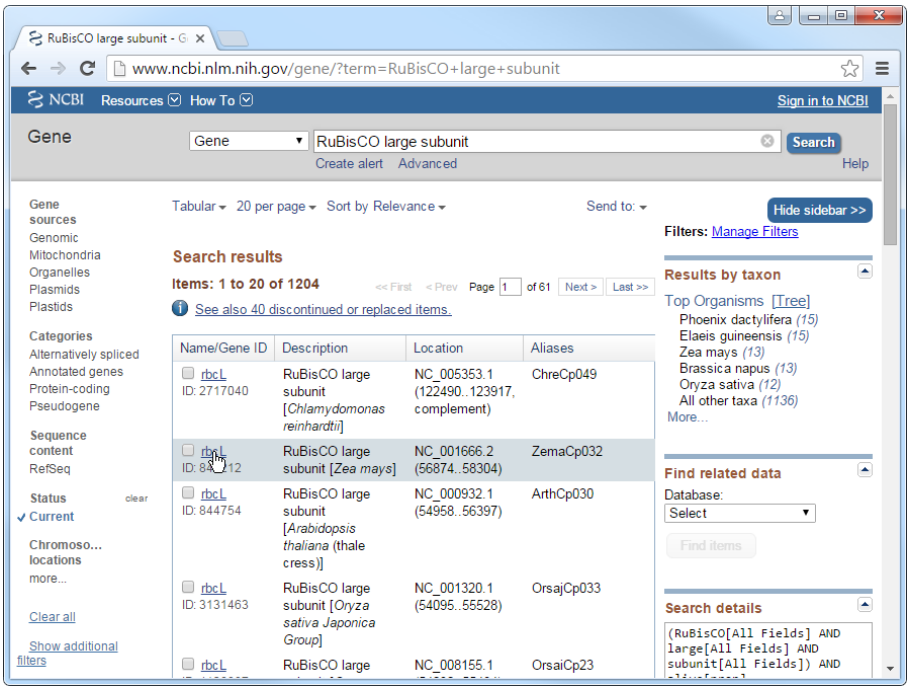


Figure 4. Select the link: **rbcL** for RuBisCO large subunit for a given species, here we will use *Zea mays*.

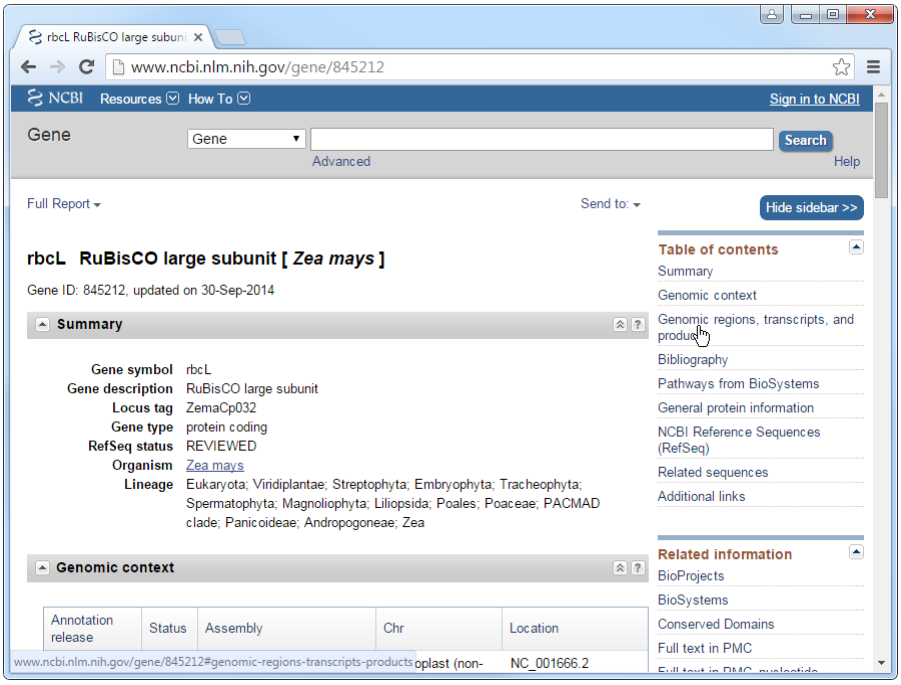
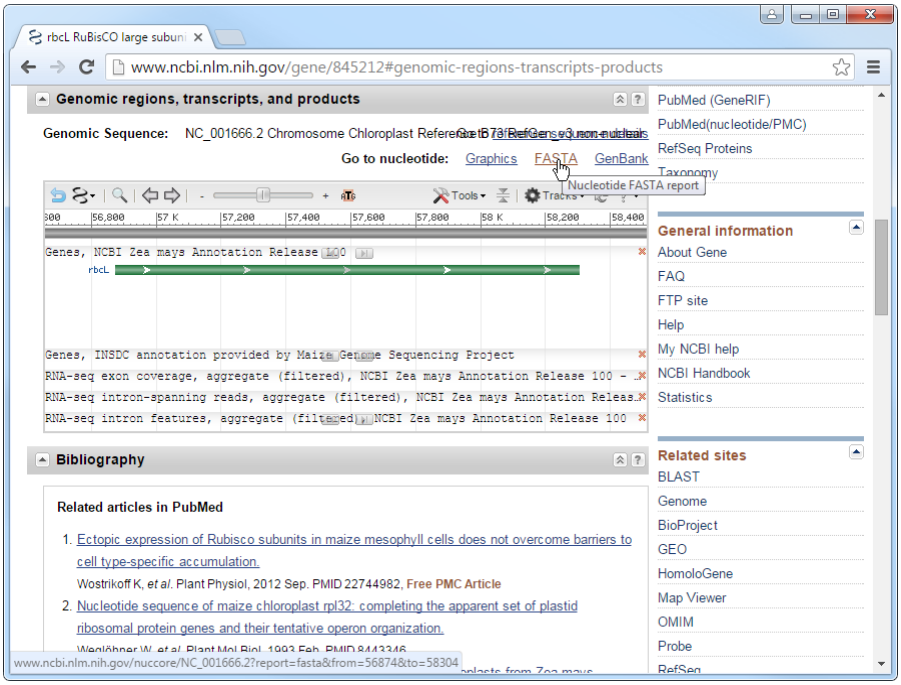
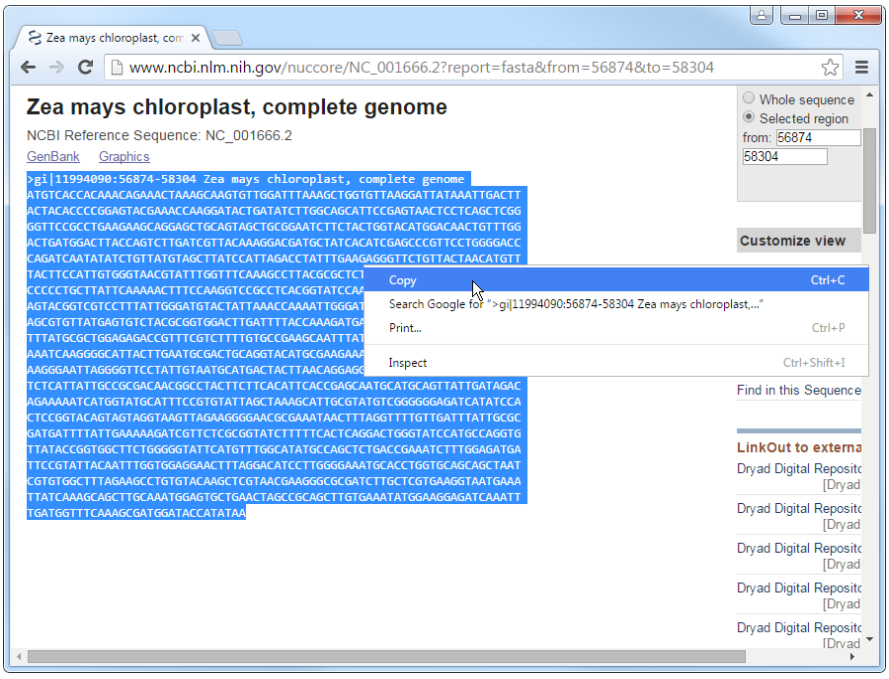


Figure 5. Click on Genomic regions, transcripts, and products in the table on contents.



81

82 Figure 6. Click FASTA



83

84 Figure 7. Copy the sequence of rcbL gene from *Zea mays*.

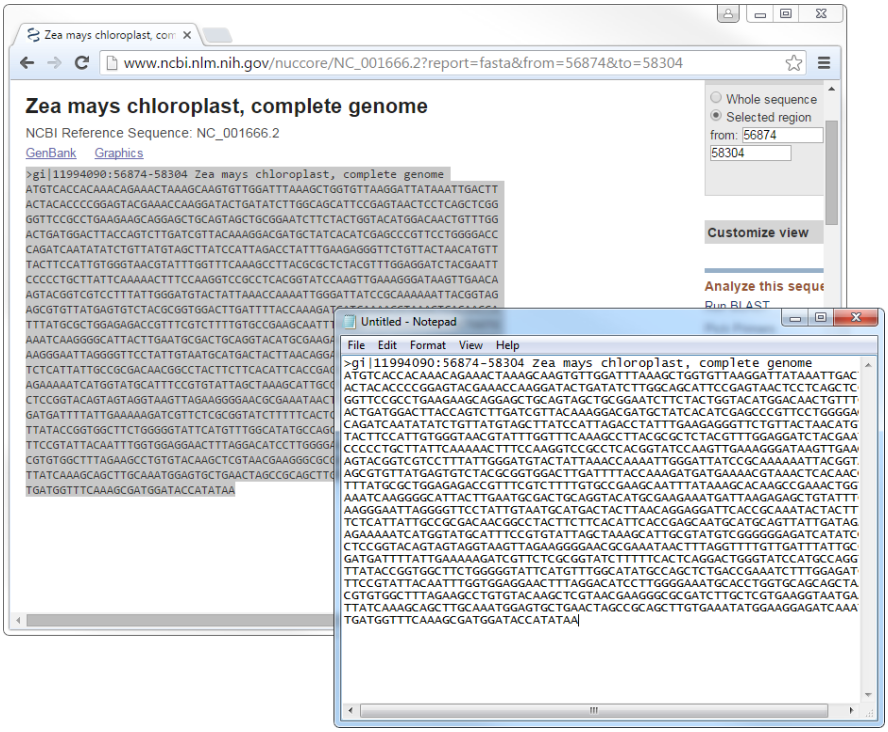


Figure 8. Paste the sequence data into a Notepad(PC) or Texteditor(Mac/Linux)

(To find Notepad on a PC with Windows go to the Start menu, All Programs then click Accessories and you should see Notepad.)

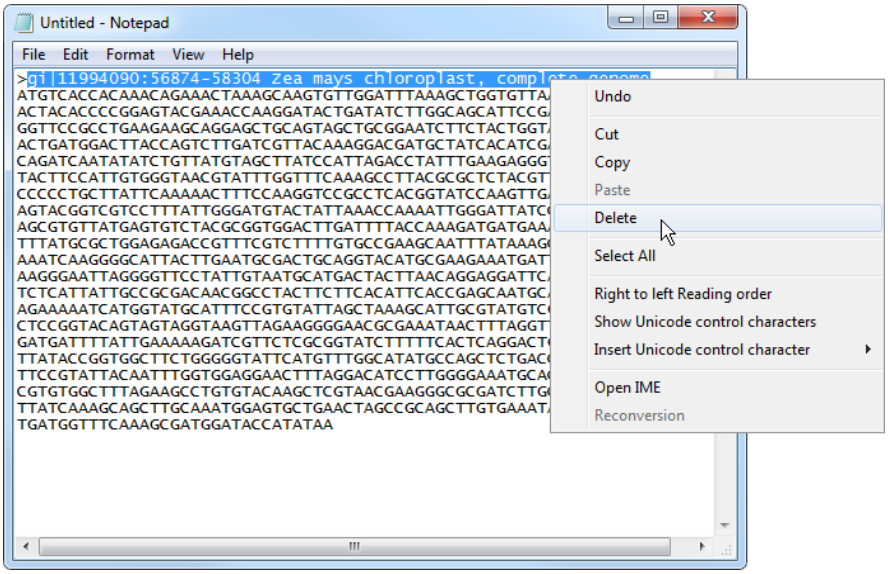


Figure 9. After pasting into *Notepad,* leave the prompt sign **>** and delete text before the DNA sequence, then replace deleted text with Corn, the common name for *Zea mays*.

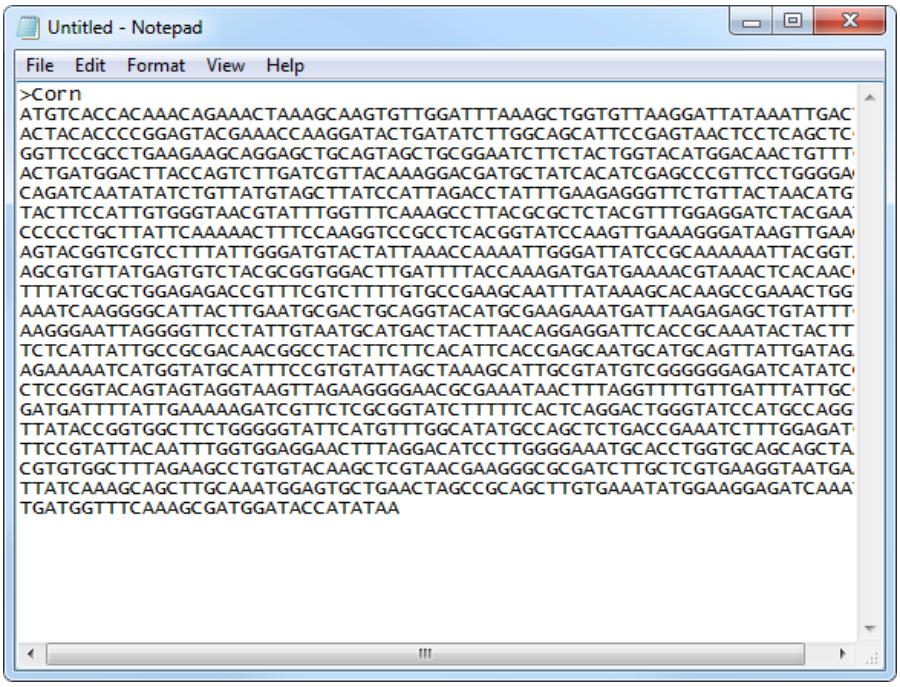


Figure 10. Notepad should appear as shown above.

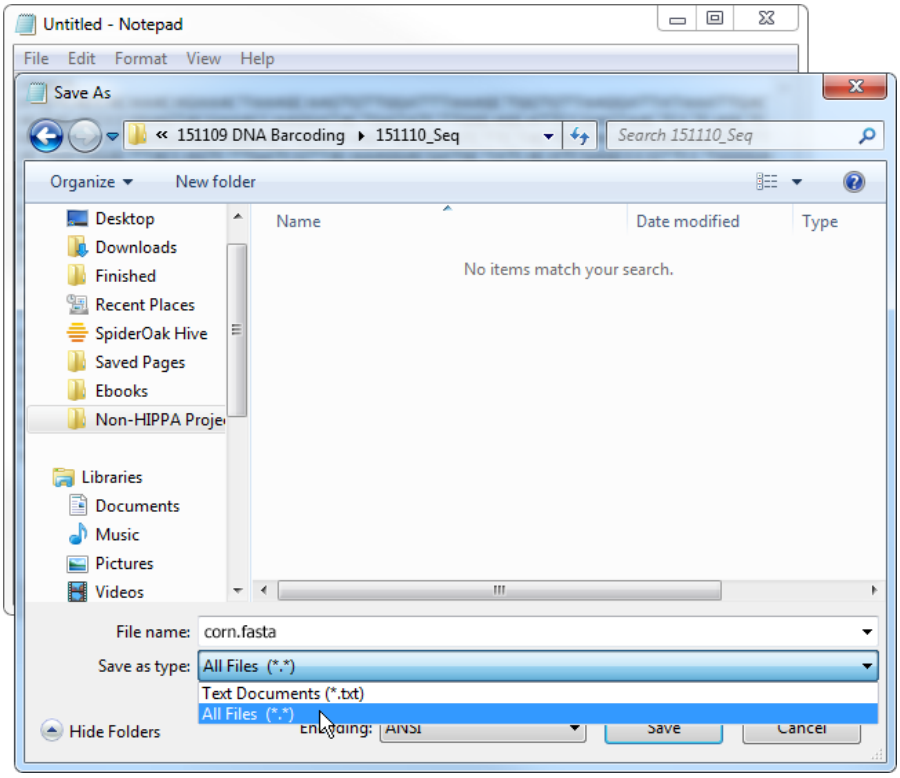


Figure 11. Save this file as corn.fasta and in the drop down choose All files (*.*).

We now have the *rbcL* gene sequence for one species. We need to collect sequences for nine other species for comparison in MEGA (See Table 1). In addition, we will use *Euglena viridis* as our outgroup. Repeat the procedure with each species listed below by typing the GenBank Gene ID into the search box shown as shown in Figure 3. Make sure to select Gene to the right of the search box when searching by the Gene ID number.

Table 1. List of plants from which the rcbL gene that can be used to create the phylogenetic tree, their GenBank ID numbers (Accession Numbers), and the suggested file names.

Common name	***GenBank ID***	***File name***
Corn	845212	corn.fasta

Thale cress	844754	thalecress.fasta
Rice	4126887	rice.fasta
Tobacco	800513	tobacco.fasta
Potato	4099985	potato.fasta
Liverwort	2702554	liverwort.fasta
Sunflower	4055709	sunflower.fasta
Grape	4025045	grape.fasta
Cucumber	3429289	cucumber.fasta
Spinach	2715621	spinach.fasta

Obtaining the Outgroup:

Euglena viridis will be used as the outgroup in our evolutionary tree. We will use the NCBI GenBank to locate the sequence for *Euglena viridis*.

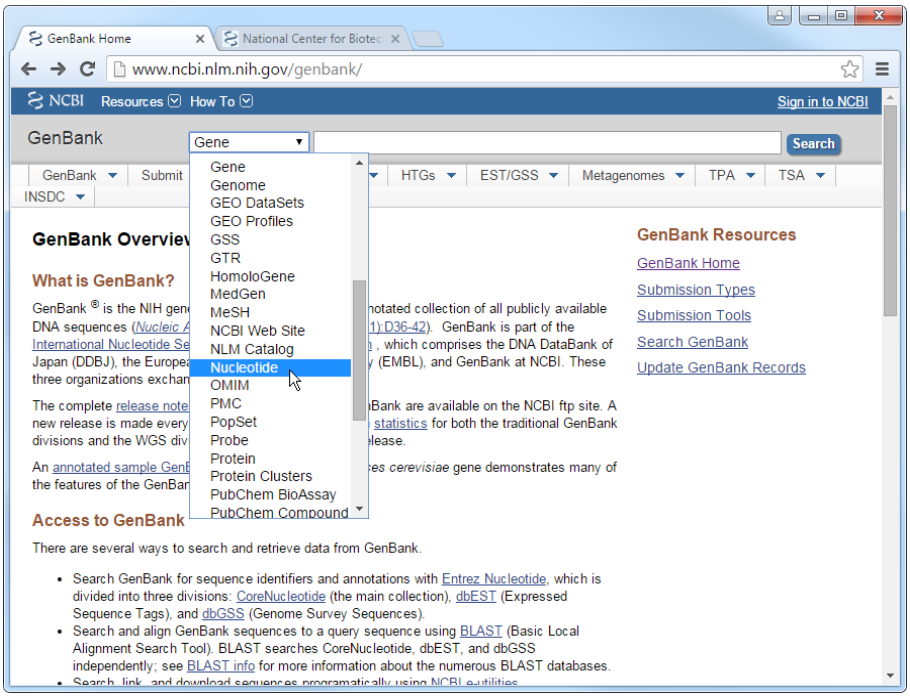


Figure 12. Starting from GenBank Home we will select the Nucleotide search filter option to the left of the search box.

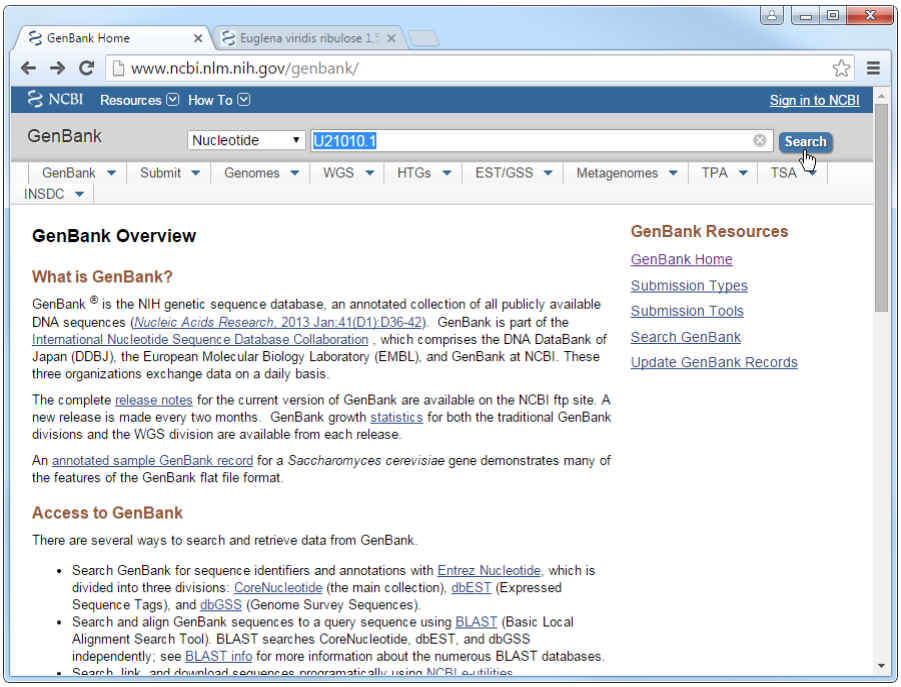
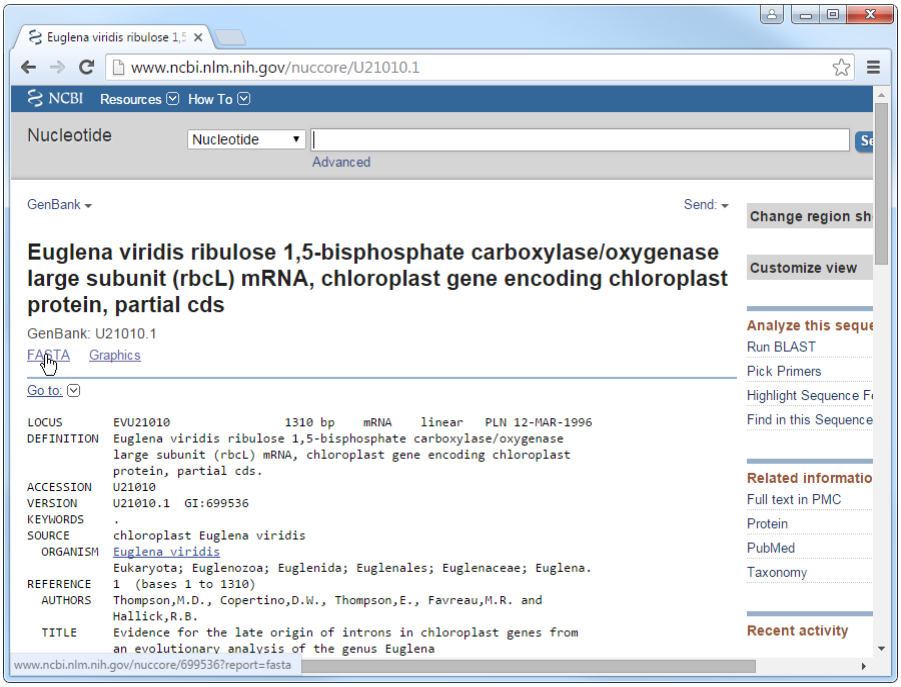
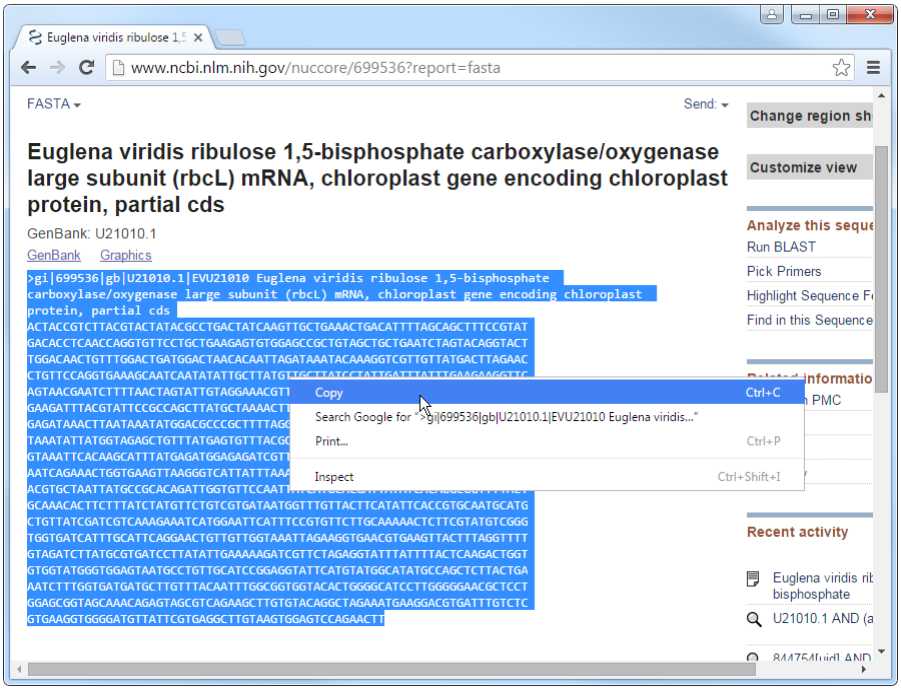


Figure 13. Type **U21010.1** into the search box and click search.



114

115 Figure 14. Click FASTA



116

117 Figure 15. Copy the sequence and pate into a new Notepad file. Repeat steps in figures

118 10 and 11. Save the file as euglenaviridis.fasta.

119 ***Building the Evolutionary Trees:***

The first step in the process of building evolutionary trees with this molecular data is to download MEGA and install it on the computers that are going to be used for the project. This tutorial features the latest stable version of MEGA at the time of print (http://www.megasoftware.net/mega.php).

Instructions for using MEGA:

1. Open MEGA.
2. Click on Align then selected Edit/Build Alignment

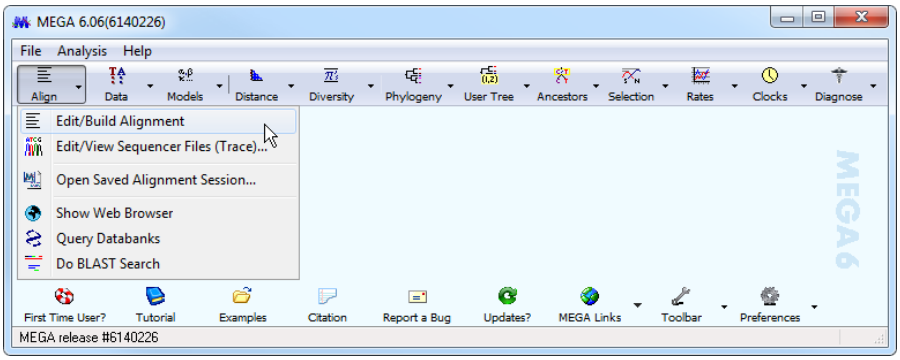


Figure 16. Click Edit/Build Alignment

3. Create a new alignment. A secondary menu will appear that requires you to select an option. Choose Create a new alignment option and click OK.

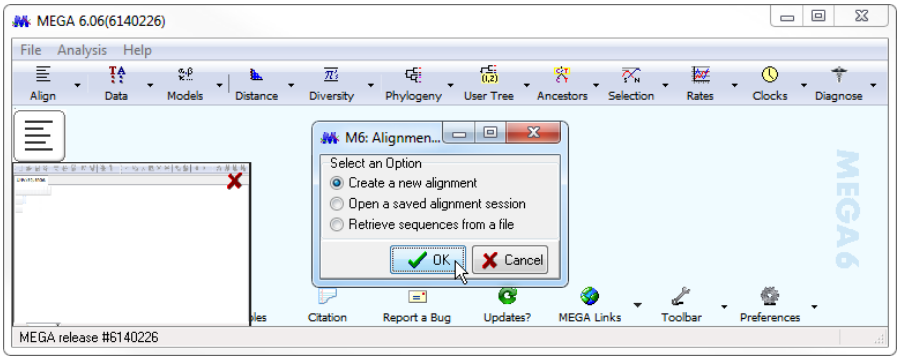


Figure 17. Select Create a New Alignment and click OK

4. A second submenu will appear asking you to select the type of sequence data that will be used to build the alignment. Select the DNA option (Figure 18). This will open the MEGA Alignment Explorer in a new window (Figure 19).

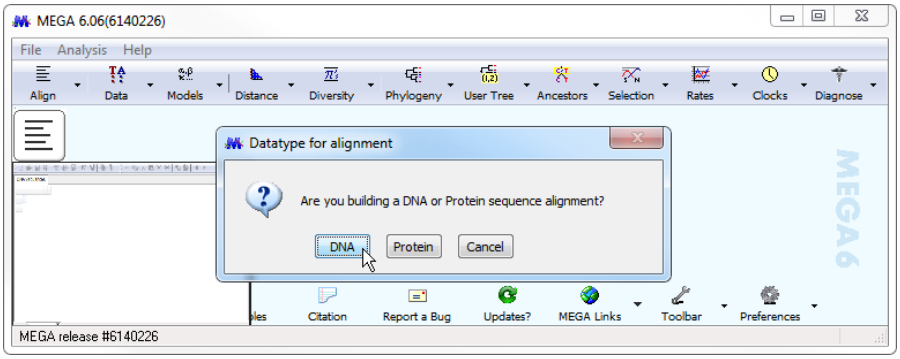


Figure 18. Click DNA

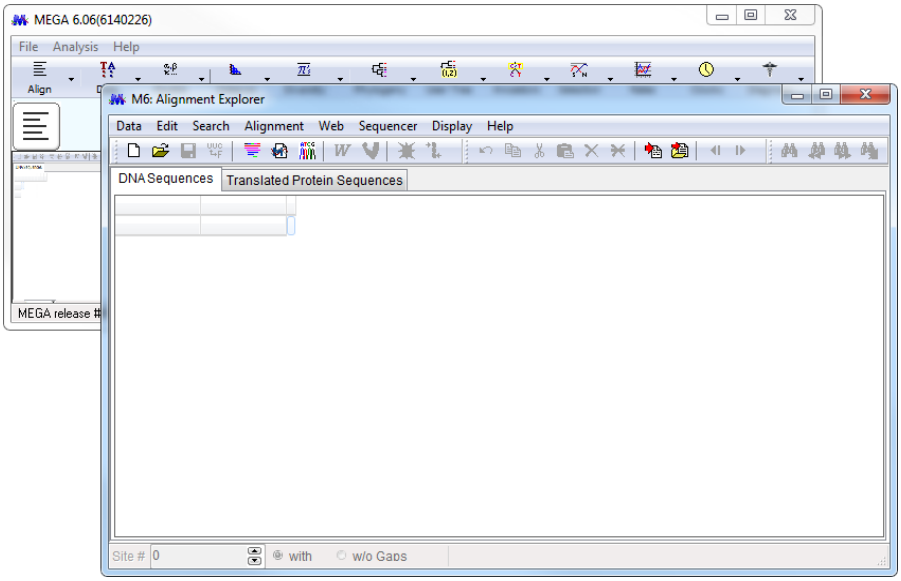
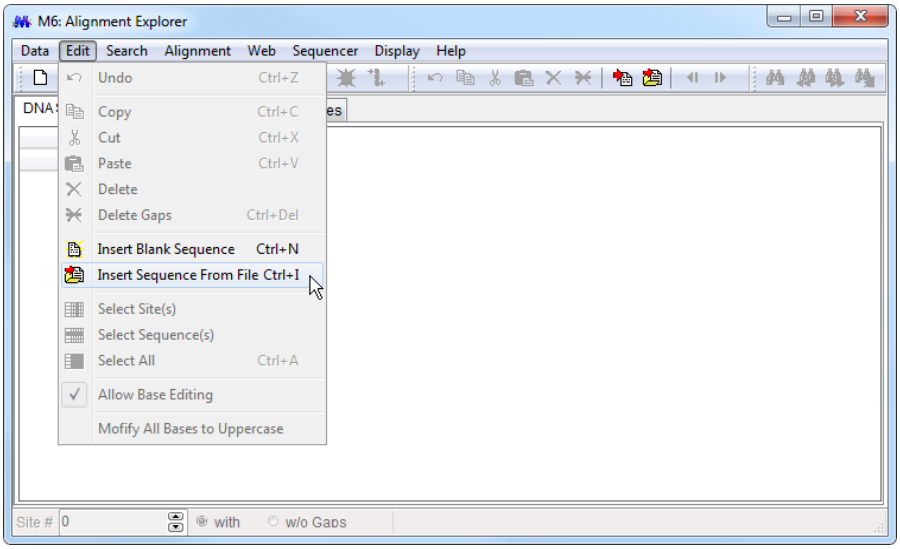


Figure 19. Alignment Explorer opens

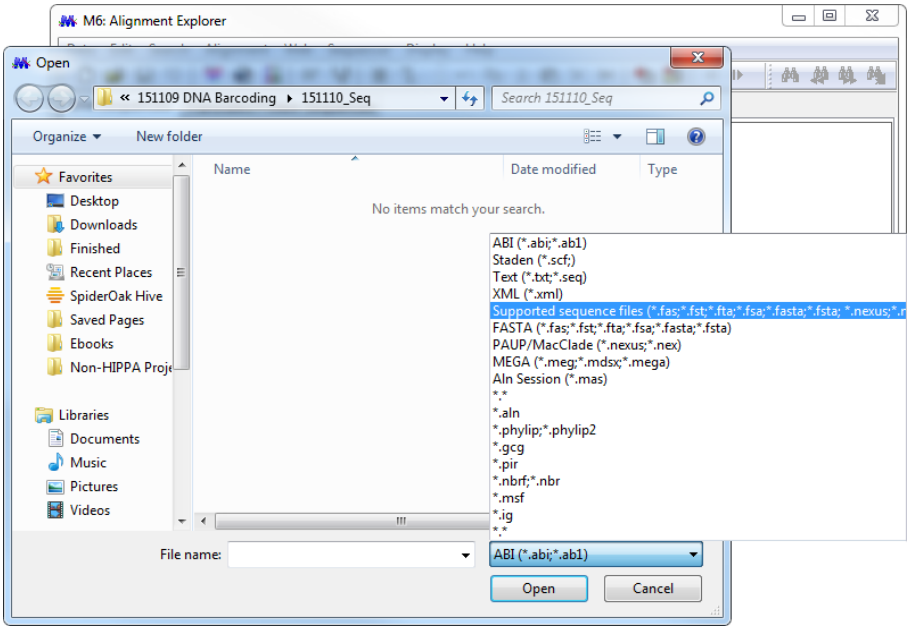
5. At the top of the MEGA Alignment Explorer Window select the Edit menu by clicking on it. From this menu, select the Insert Sequence From File option (Figure 20). This will open a new window.



143

144 Figure 20. Click Insert Sequence from File

145 6. In the window of the Open, select the **.**fasta data files saved earlier



146

147 Figure 21. Select Supported sequence files

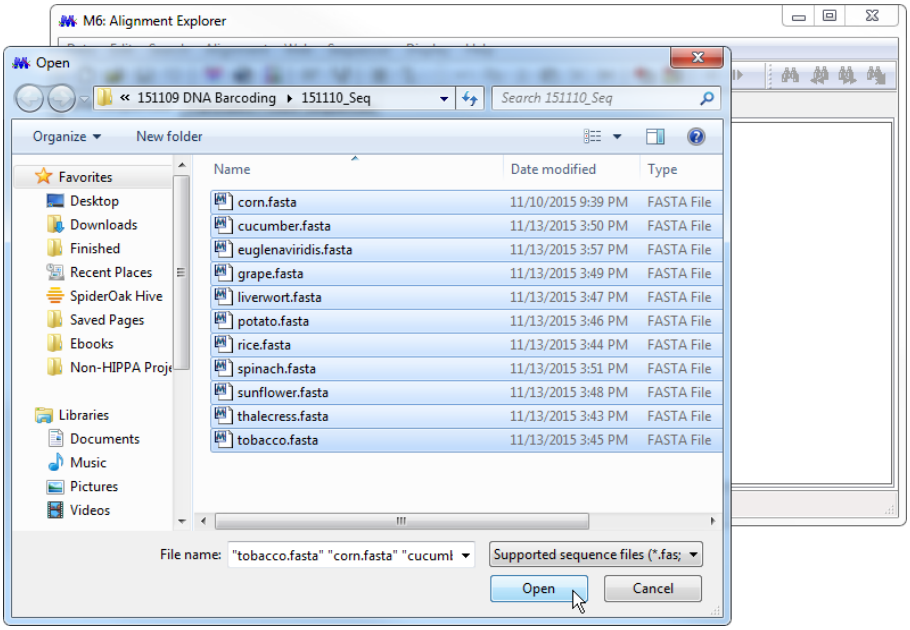


Figure 22. Select the .fasta files and click Open

7. Once you have selected all of the files that you wish to upload, select the Open button by clicking on it.

8. Once the sequences are loaded into MEGA, we want to align them. This is done by going to the Alignment menu at the top of the Alignment Explorer window. Click to open the dropdown menu and select Align by ClustalW by clicking on it (Figure 23).

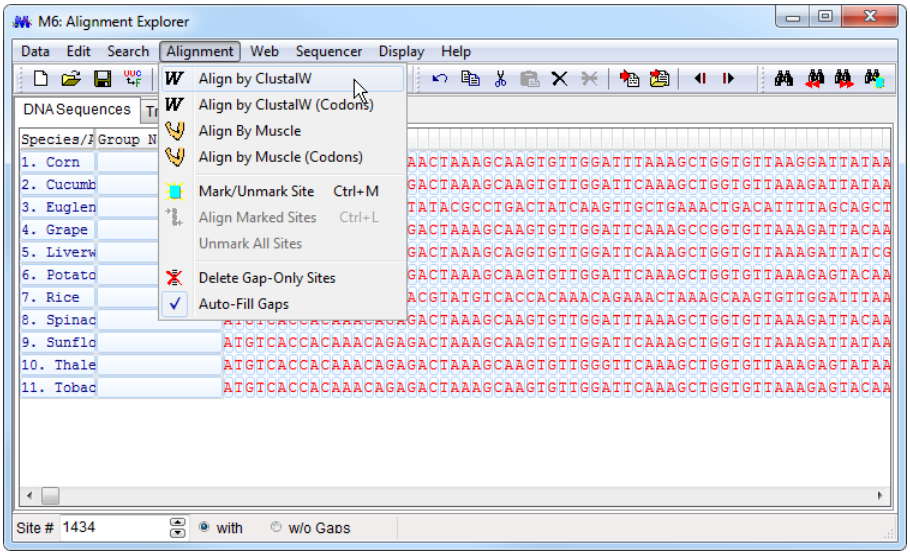


Figure 23. Align By ClustalW

9. This will open another window that is filled with ClustalW parameters. For our purposes, the default settings are adequate. Select the OK option at the bottom of this menu to proceed (Figure 24). This will set in motion the alignment algorithm. Aligning the sequences may take several minutes depending on the size and number of the sequences being examined.

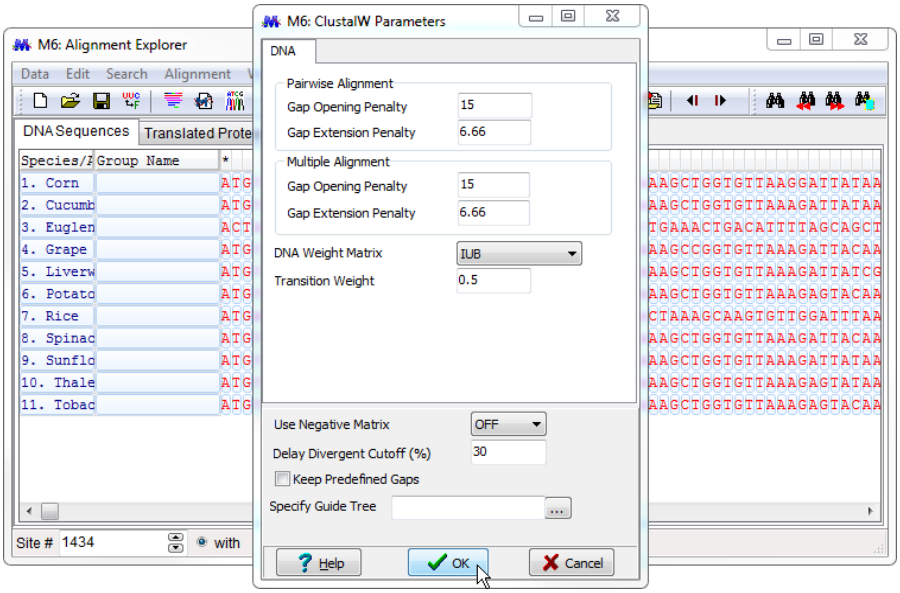


Figure 24. ClustalW Parameters Dialog. Leave all set to their defaults.

10. Now that the sequences are aligned (Figure 10), we may need to trim the ends of the sequences so that they line up nicely and do not contain too much excessive data. To do this, look for an asterisk (*) in the line above the first sequence in the Alignment Explorer window. At the first asterisk from the beginning of the sequences, use Click and Drag to select the nucleotides from there to the beginning of the sequences to be removed. (Figure 25).

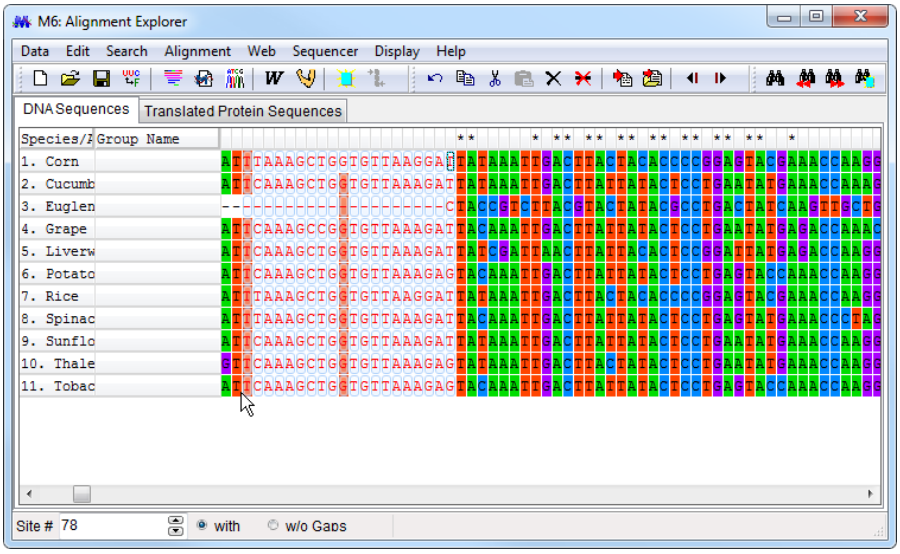


Figure 25. Remove the Excess Nucleotides

11. This process will need to be repeated for the tail ends of the sequences. Here find the last asterisk from the end, then Click and Drag to highlight the nucleotides to the end of the sequences.

12. Now the sequences are aligned and trimmed (Figure 26).

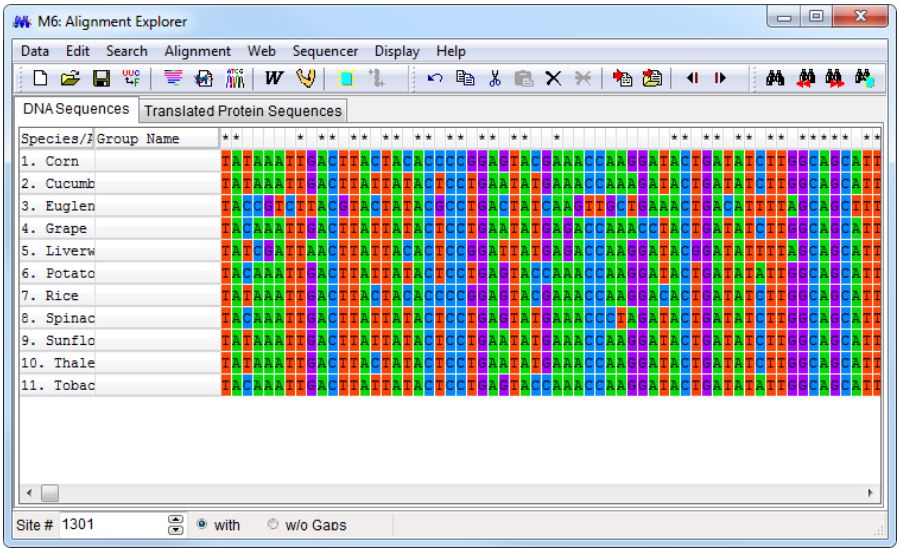


Figure 26. Sequences are aligned and trimmed

13. Now, we need to save the Alignment Session so that the data is saved in a format that MEGA can use to build the phylogenetic trees. Save the Alignment Session by selecting the Data menu at the top right of the Alignment Explorer window. Click on the Save Alignment option. (Figure 27). After this has been completed and the session saved, close the Alignment Explorer window.

Figure 27. Save Session

14. Now, we need to open the saved alignment session by selecting the File menu in the general MEGA window and clicking on the Open a File/Session option (Figure 28).

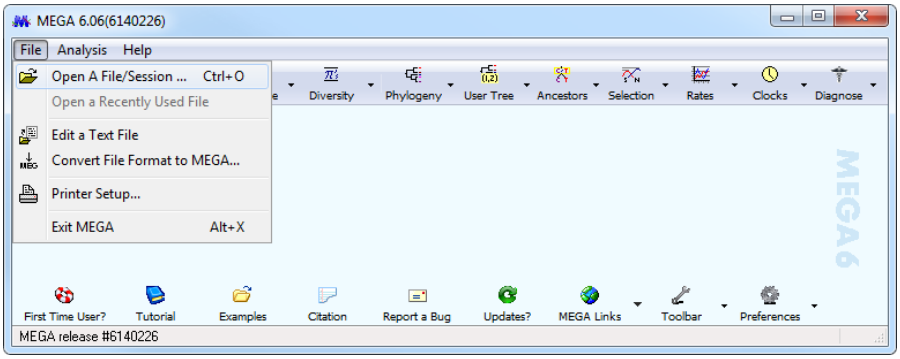


Figure 28. Open A File-Session

15. Select the saved alignment file (it will be a .MAS file) in the Open a File window and select Open by clicking on it (Figure 29).

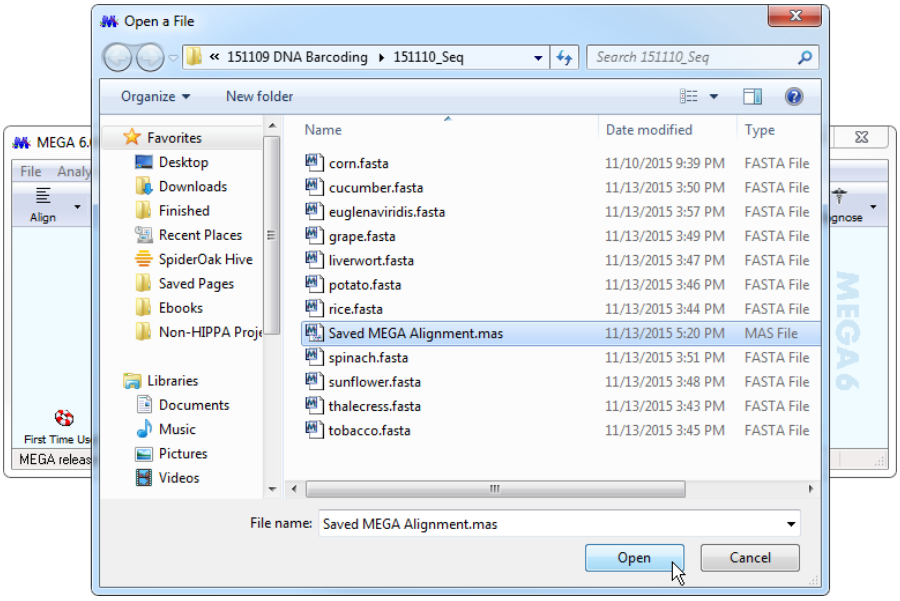


Figure 29. Click Open

16. This will bring up a second window where you have the choice to open the .MAS to analyze it or align it. Select Analyze by clicking on it (Figure 30).

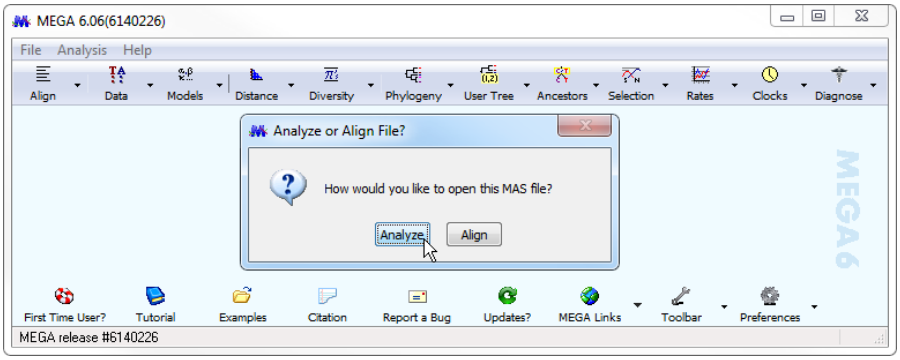


Figure 30. Click Analyze

17. To construct a phylogenetic tree, select the Phylogeny menu mid-way through the second menu bar at the top of the MEGA window. Here we will continue our example

with a Neighbor-Joining tree, but the process is the same for other types of phylogenetic trees. Select Construct/Test Neighbor-Joining Tree (Figure 31).

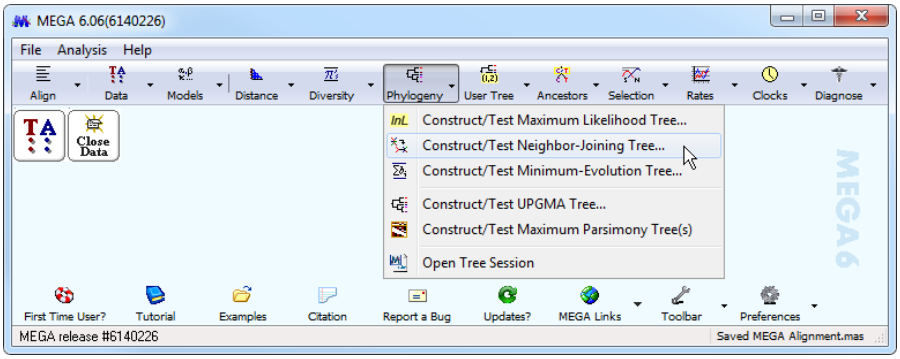


Figure 31. Construct-Test Neighbor-Joining Tree

18. A menu will appear that asks if you want to use the currently active data sheet, select Yes (Figure 32).

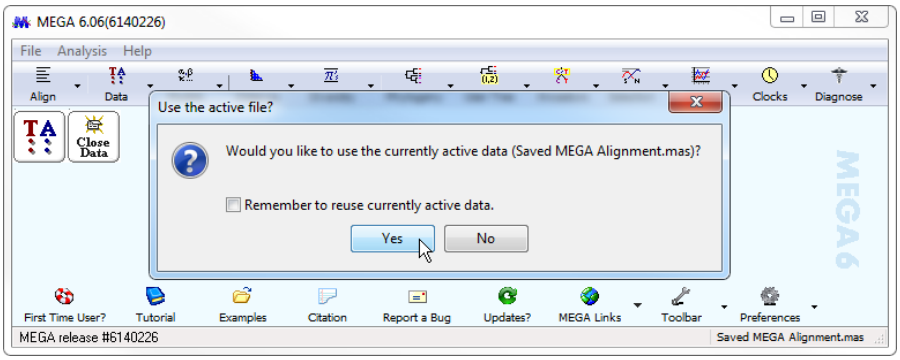


Figure 32. Click Yes

19. This will open a dialogue box called Analysis Preferences. For the Statistical Method select Neighbor-Joining and for the Test of Phylogeny select Bootstrap Method. In the field entitled No. of Bootstrap Replications, select 1000 to obtain stable estimates of reliability of the tree. For the Substitution Type start by selecting Nucleotide followed by selecting the Jukes-Cantor Model. Here all other fields are left at their default values. To generate the tree, click on Compute (Figure 33).

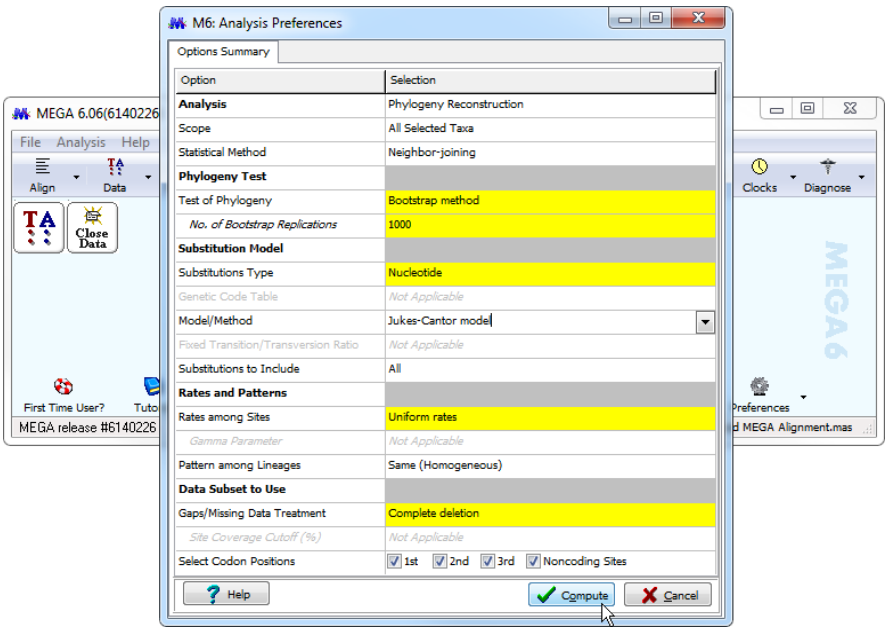


Figure 33. Click Compute.

20. After a few minutes, a tree will be generated. The length of time that this takes will depend in part on the length and number of sequences that are being used to create the tree (Figure 34).
21. The numbers on the branches of the tree represents the Bootstrap value, which is the statistical support that each branch receives by the Bootstrap analysis. Higher numbers mean that the branch has higher support and is most likely to be a real branch (Figure 34).

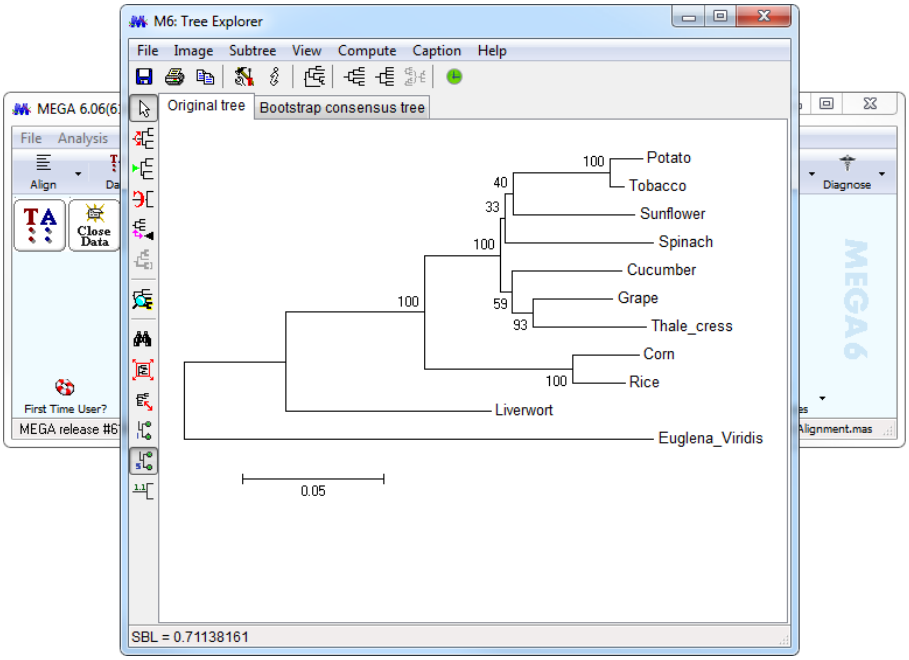


Figure 34. Generated Tree

22. To simplify the tree, we now want to condense or cut out the branches that have less support and are less likely to be true branches. To do this, go to the Compute menu on the TreeExplorer menu and select Condensed Tree (Figure 35).

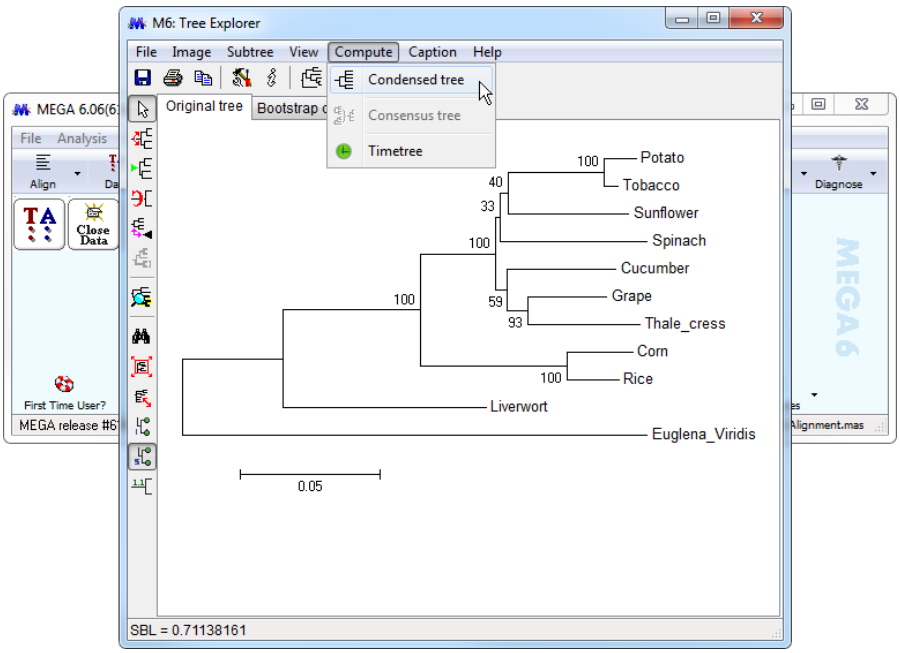


Figure 35. Select Condensed Tree

23. This opens a new menu, Tree Options. Select the Cutoff submenu and input 50 for the Cut-off Value for Condensed Tree, then click OK (Figure 36). Leave all other values at their defaults.

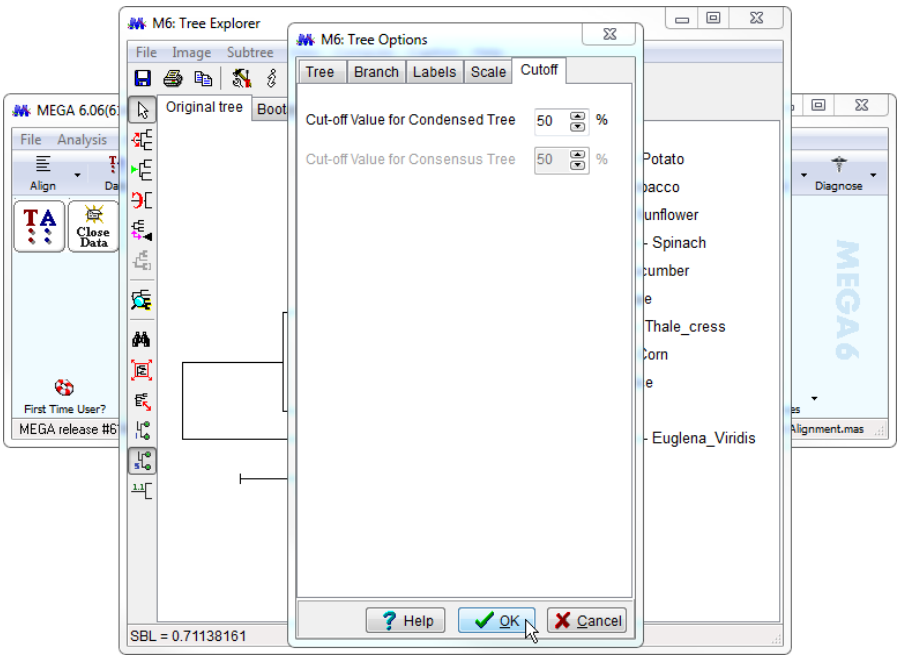


Figure 36. Click OK.

24. Now, the tree in the TreeExplorer will reflect the changes. All branches that have less than 50% support will have been removed. (See Figure 37)

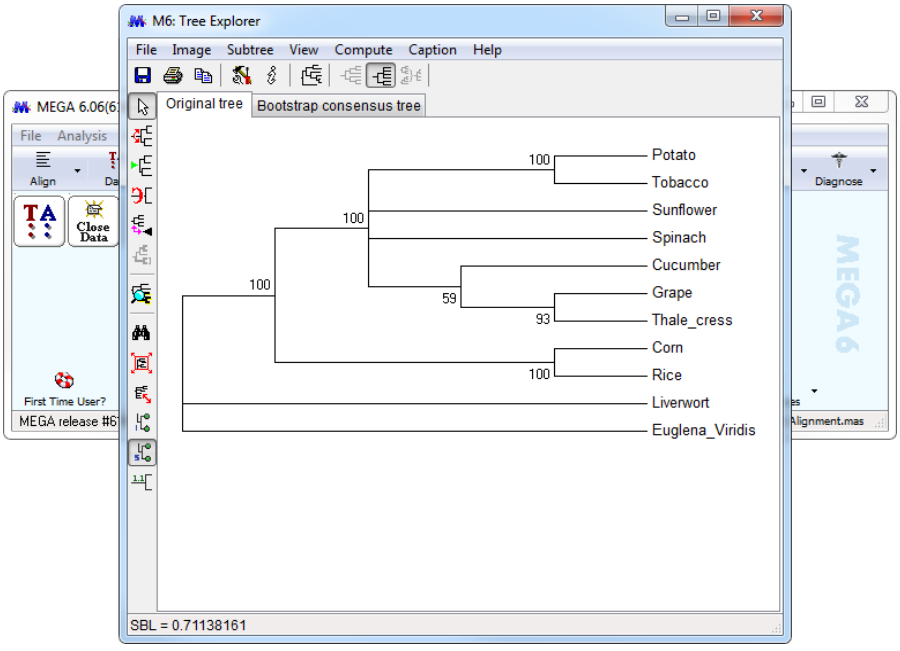


Figure 37. Condensed Tree

25. The last thing that we need to do is to set our out group. In our example here it is Euglena viridis. To do this, right click on the branch that has Euglena viridis. This brings up a submenu. In this submenu, select the Place Root option. This provides us with a rooted phylogenetic tree. (See Figure 38)

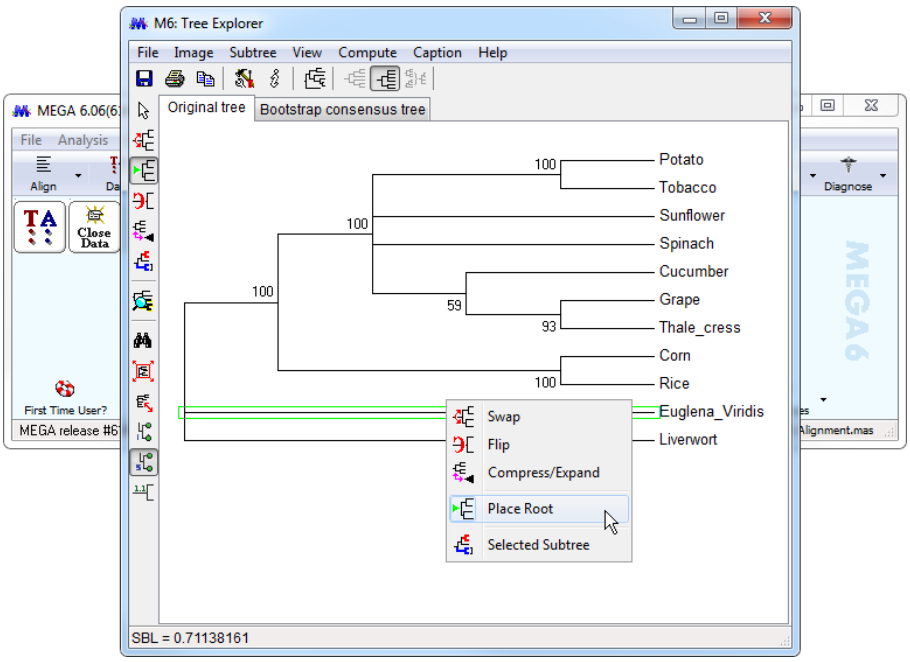


Figure 38. Place Root

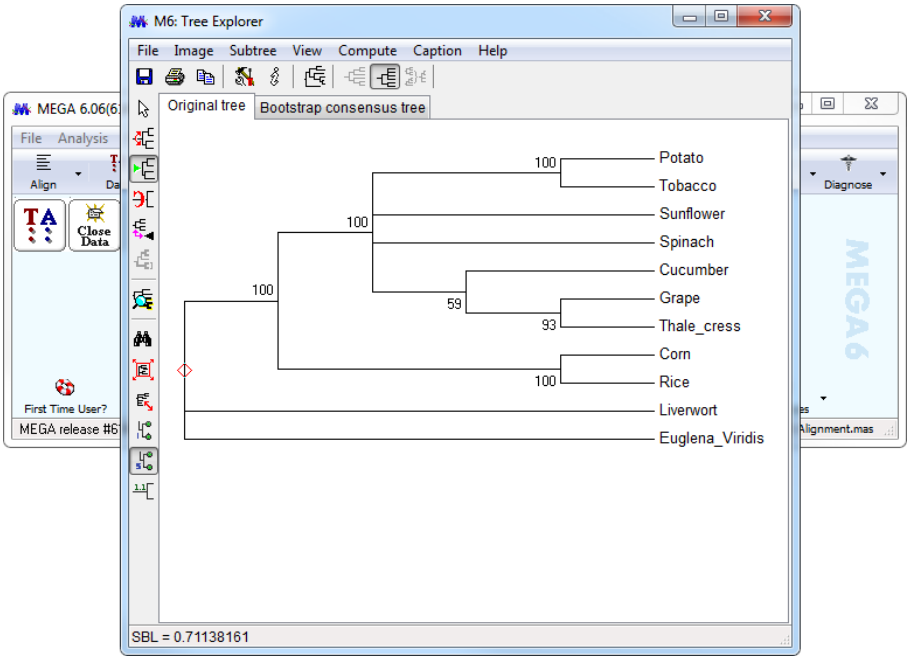


Figure 39. *Euglena viridis* as the Root

The tree can be saved as a PDF or printed out. To save the tree as a PDF, go to the Image menu and select Export as PDF. A window will pop up and you can save the file there. To Print the tree, there is a Printer Icon that you can use just below the upper menu.

Assessment Questions

In order to ensure that students understand the output of this exercise, questions along the following lines may be asked: Which species are most closely related? Give an example of sister taxa. Why do you think that corn and rice are so closely related? How many base pairs were included in your analysis? How do you think the length of the gene sequence used affects the validity/reliability of your results? Why does the liverwort group with the outgroup? These questions, or similar, will let the instructor assess not only if the student has understood the process that they have gone through to create the phylogenetic tree, but in conjunction with

the questions that the students should consider as they build the tree, the student's understanding of the process.

Additional Background Information:

The NJ and ML methods for building evolutionary trees rely on different statistical principles (Nei and Kumar 2000; Tamura et al. 2011). In NJ, the least squares method is used along with pairwise evolutionary distances (Nei and Kumar 2000). In ML, the maximum likelihood is optimized such that the inferred tree is the most likely tree (Nei and Kumar 2000). Generally, they will produce very similar results, but NJ is much faster. Despite slight differences in the branching patterns between NJ and ML trees, they both are robust methods for building evolutionary trees. The Jukes-Cantor model is simply a mathematical model that describes the change of one the nucleotides in the DNA sequence to another one, over time (Nei and Kumar 2000). All of its parameters are automatically estimated by MEGA.

Both NJ and ML produce trees that are unrooted, even though they are frequently drawn from left to right. In this case, if one knows the outgroup then it can be used to properly root the tree. Choosing a proper out group can be a difficult task and may require some trial and error (Nei and Kumar 2000). A good out group should be similar to the sequences in question, but different enough so that the computer program can see the differences (Nei and Kumar 2000).

Acknowledgements:

The authors acknowledge the editors and Dr. Sudhir Kumar, Director of iGEM (Institute for Genomics and Evolutionary Medicine at Temple University for their thoughtful assistance with this article.

References:

279 Hall BG (2013) Building phylogenetic trees from molecular data with MEGA. Mol. Biol. Evol.
280 30:1229-1235.
281 Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA 5L: Molecular
282 evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum
283 parsimony methods. Mol. Biol. Evol. 28:2731-2739
284 Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S (2013) MEGA6: Molecular evolutionary
285 genetics analysis version 6.0. Mol. Biol. Evol. 30:2725-2729
286 Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, Oxford,
287 UK. pp 333.
288 Newmaster, SG, Fazekas, JA, and Ragupathy S (2006) DNA barcoding in land plants: evaluation
289 rbcL in a mulitigene tiered approach. Can. J. Bot. 84:355-341.
290 Ryan, MP, Adley, CC, Pembroke, JT (2013) The use of MEGA as an educational tool for
291 examining the phylogeny of antibiotic resistance genes. In. Microbial pathogens and strategies
292 for combating them: science, technology and education. Ed. A Menendez-Vilas. pp. 736-743.