TIPS, TRICKS & TECHNIQUES

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

LUCAS NEWMAN, AMANDA L. J. DUFFUS, CATHY LEE

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 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 the data into a text editor, import the data into MEGA, and use the dataset to create phylogenetic trees.*

Key Words: *MEGA; evolutionary trees; molecular data; neighbor joining; maximum likelihood.*

○ Introduction

Phylogenetics is the study of the evolutionary relatedness between different groups of organisms (Nei & Kumar, 2000). These groups can be at 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, in which diffrent 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.

Phylogenetics is the study of the evolutionary relatedness between different groups of organisms.

Here, we show students how to build evolutionary trees using the MEGA software package (version 6; http://www.megasoftware.net) and thereby introduce them to two of the most commonly used methods for inferring evolutionary relationships among species by using gene sequences: neighbor joining (NJ) and maximum likelihood (ML). According to its authors, MEGA is frequently used in educational settings in advanced classes (Sudhir Kumar, personal communication; Ryan et al., 2013). However, many K–12 instructors are not familiar with its potential to introduce the concepts of evolutionary biology to students in a hands-on, discovery-based pedagogy using gene sequences. There are multiple online resources that provide such gene sequences for a multitude of species (e.g., GenBank, which is available from the National Center for Biotechnology Information [NCBI]; Hall 2013). Both DNA and protein sequences are available, and several informative tutorials are provided on how to use these on the NCBI website. Literally unlimited sequence data from thousands of genes from animals, plants, protists, bacteria, and viruses are available through GenBank.

○ Overview

Project Goal

We will build an evolutionary tree using the *rbcL* gene sequence, which is commonly used to study the evolutionary relationships between plants (see Newmaster et al. 2006). Sequence data pertaining to the *rbcL* genes from many plant species are available through GenBank, and we will gather our dataset from this resource. The *rbcL* gene sequence data will then be imported into MEGA, aligned, and used to build a phylogenetic tree.

Helpful Prior Knowledge & Potential Context

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

The American Biology Teacher, Vol. 78, No 7, pages. 608–612, ISSN 0002-7685, electronic ISSN 1938-4211. DOI: 10.1525/abt.2016.78.7.608.

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 the freely available software MEGA (http://www.megasoftware.net/mega.php).
3. Understand the meaning of ancestral vs. recent species, understand clades, and be able to interpret evolutionary relationships among species.

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

1. Why was the *rbcL* gene used?
2. Which organelle does the *rbcL* gene originate from?
3. What function does the protein product of the *rbcL* 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).

○ Getting Started

1. Google search for GenBank and click on the result for GenBank Home (http://www.ncbi.nlm.nih.gov/genbank/).
2. Use the dropdown next to the word GenBank to change from default Nucleotide and select Gene.
3. Type **RuBisCO large subunit** in the entry box to the right of dropdown and click Search.
4. Select the link *rbcL* for the RuBisCO large subunit for a given species – here, we will use *Zea mays*.
5. Click on Genomic regions, transcripts, and products in the table of contents.
6. Click FASTA
7. Copy and paste the sequence of *rbcL* gene from *Zea mays* (Figure 1).
8. 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*.
9. Click File then Save. A Save As dialog box will appear. In the Save as type: drop down choose All files (*.*) and save this file as "corn.fasta".
10. We now have the rbcL gene sequence for one species. We need to collect sequences for nine other species for comparison

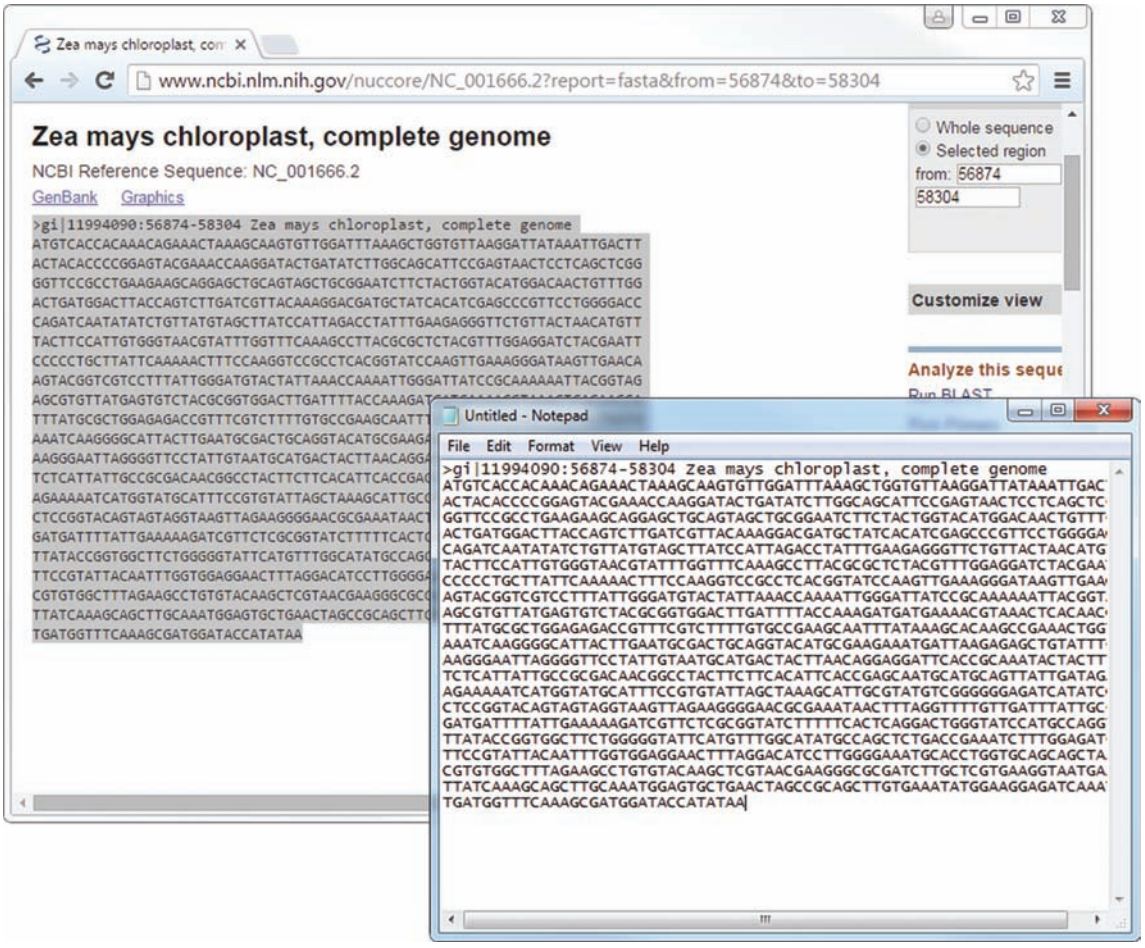

Figure 1. Paste the sequence data into a Notepad (PC) or Texteditor (Mac/Linux). To find Notepad on a PC with Windows, go to the Start Menu, All Programs, then click Accessories and you should see Notepad.

Table 1. List of plants from which the *rbcl* gene can be used to create the phylogenetic tree, with their GenBank ID numbers (accession numbers) and 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

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. Save each sequence in a separate file as suggested in Table 1.

○ Obtaining the Outgroup

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

1. Starting from GenBank Home, we will select the Nucleotide search filter option to the left of the search box.
2. Type **U21010.1** into the search box and click Search.
3. Click FASTA.
4. Copy the sequence and paste into a new Notepad file. Remember to delete everything just before the gene sequence and replace with the **>** prompt followed by the organism's name.
5. Save the file as euglenaviridis.fasta.

○ Building the Evolutionary Trees

The first step in the process of building evolutionary trees with these molecular data is to download MEGA and install it on the computers that are going to be used for the project. This tutorial features MEGA 6, the latest stable version during the creation of this guide. MEGA 7, in beta at the time of writing, will contain many of the same features in addition to various improvements over the previous versions of MEGA.

Instructions for Using MEGA

1. Open MEGA.
2. Click on Align, then select 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.
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. This will open the MEGA Alignment Explorer in a new window.
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. This will open a new window.
6. In the Open dialog box window, navigate to the directory containing the saved .fasta data files. Once in the correct folder, if no files appear, click on the File Type drop down box next to the File Name box. Select the Supported sequence files option. The data files should now be visible.
7. Select the .fasta files and click Open. Once you have selected all of the files that you wish to upload, select the Open button. 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.
8. 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. This will set the alignment algorithm in motion. Aligning the sequences may take several minutes, depending on the size and number of the sequences being examined. ClustalW Parameters dialog: Leave all set to their defaults.
9. Now that the sequences are aligned, we may need to trim the ends of the sequences so that they line up nicely and don't contain 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 2).
10. 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.
11. Now the sequences are aligned and trimmed.
12. Now, we need to save the Alignment Session so that the data are 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 Session option. After this has been completed and the session saved, close the Alignment Explorer window.
13. 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.
14. Select the saved alignment file (it will be a .MAS file) in the Open a File window and select Open by clicking on it.
15. 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. Click OK.

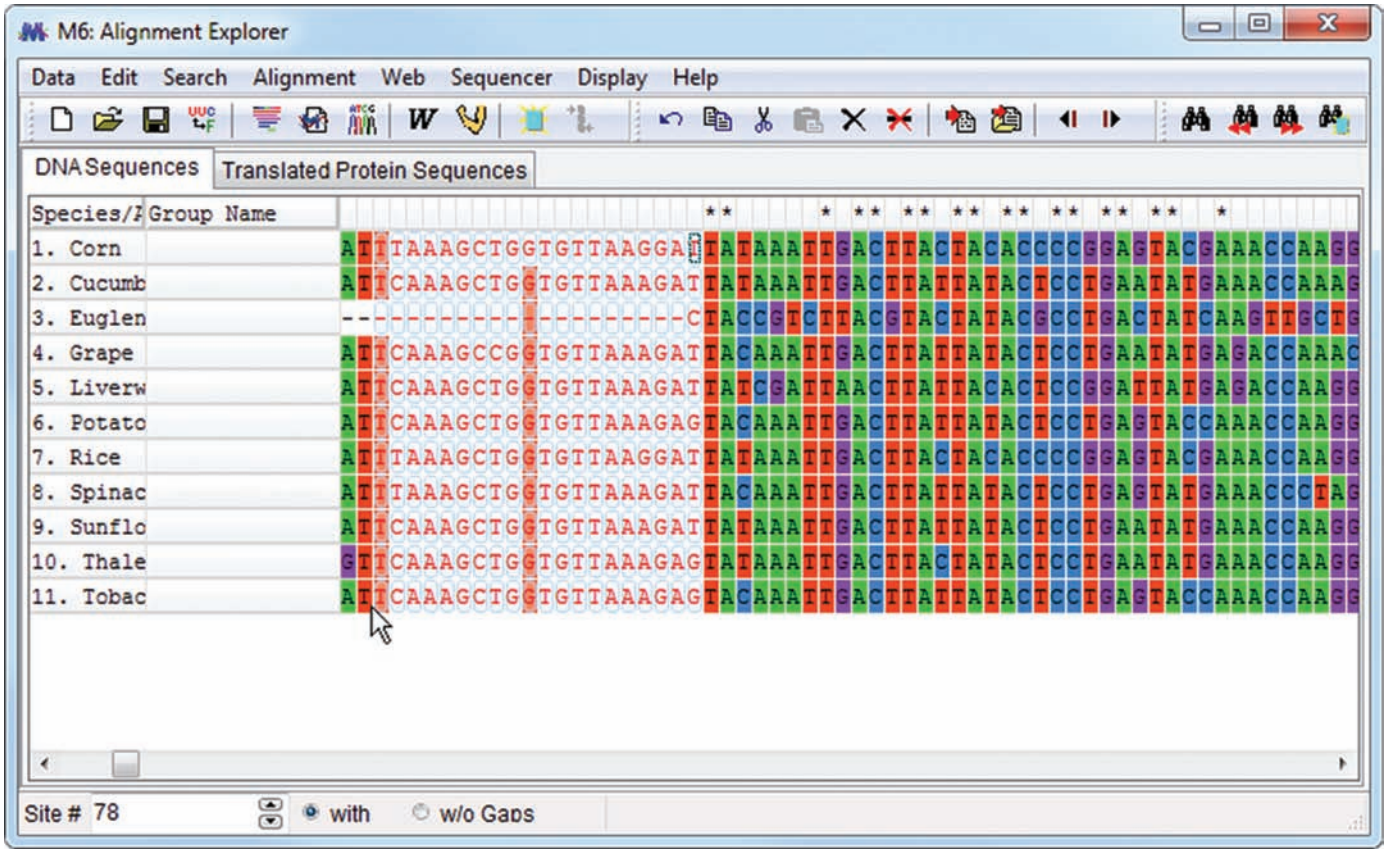

Figure 2. Remove the excess nucleotides.

16. To construct a phylogenetic tree, select the Phylogeny menu midway 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.
17. A menu will appear that asks if you want to use the currently active data sheet; select Yes. This will open a dialog 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 and then select the Jukes-Cantor Model. All other fields are left at their default values. To generate the tree, click on Compute.
18. After a few minutes, a tree will be generated. The length of time this takes will depend, in part, on the length and number of sequences that are being used to create the tree.
19. The numbers on the branches of the tree represent 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 more likely to be a real branch.
20. 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.
21. 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. Leave all other values at their defaults.
22. Now, the tree in the TreeExplorer will reflect the changes. All branches that have less than 50% support will have been removed.
23. The last thing that we need to do is set our outgroup. In our example here, it is *E. viridis*. To do this, right click on the branch that has *E. viridis*. This brings up a submenu. In this submenu, select the Place Root option. This provides us with a rooted phylogenetic tree.

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 Save as PDF file. A window will pop up and you can save the file there. To print the tree, there is a Printer icon that you can click just below the upper menu.

○ Assessment Questions

To ensure that students understand the output of this exercise, questions along the following lines can 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 liverwort group with the outgroup? These questions, or

similar ones, will let the instructor assess not only whether 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 & Kumar, 2000; Tamura et al., 2011). In NJ, the least squares method is used along with pairwise evolutionary distances (Nei & Kumar, 2000). In ML, the maximum likelihood is optimized such that the inferred tree is the most likely tree (Nei & 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 of the nucleotides in the DNA sequence to another one, over time (Nei & 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 outgroup can be a difficult task and may require some trial and error; a good outgroup should be similar to the sequences in question, but different enough that the computer program can see the differences (Nei & Kumar, 2000).

Acknowledgments

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

Web Links

YouTube instructions (https://youtu.be/uMy3HB94EVw) and the manuscript with detailed figures are found on the website (http://faculty.gordonstate.edu/clee/Accepted-%20151104%20MEGA%20-%20How%20to%20do%20it%20-%20NATB%20-%2012-1-2015.pdf).

References

Hall, B.G. (2013). Building phylogenetic trees from molecular data with MEGA. *Molecular Biology and Evolution, 30,* 1229–1235.

Nei, M. & Kumar, S. (2000). *Molecular Evolution and Phylogenetics.* Oxford University Press, Oxford, UK. p. 333.

Newmaster, S.G., Fazekas, A.J. & Ragupathy, S. (2006). DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. *Canadian Journal of Botany, 84,* 335–341.

Ryan, M.P., Adley, C.C. & Pembroke, J.T. (2013) The use of MEGA as an educational tool for examining the phylogeny of antibiotic resistance genes. In A. Méndez-Vilas (Ed.), *Microbial Pathogens and Strategies for Combating Them: Science, Technology and Education* (pp. 736–743). Badajoz, Spain: Formatex Research Center.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution, 28,* 2731–2739.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution, 30,* 2725–2729.

LUCAS NEWMAN (newmalu@ohsu.edu) is in the Department of Medical Informatics and Clinical Epidemiology, School of Medicine, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd., Portland, OR 97239. AMANDA L. J. DUFFUS and CATHY LEE (clee@gordonstate.edu) are in the Department of Biology, Gordon State College, Barnesville, GA 30204.